Statistical Methods

1.0 Principle, Spirit, and Intent

Statistical data analysis should be used in conjunction with qualitative descriptions in casework. Statistical procedures, calculations and tests should be undertaken in an appropriate and organized manner, and should be documented to enable interpretation, replication and verification by independent parties.

2.0 Purpose and Scope

Statistics are used in forensic anthropology to inform scientific inferences via the collection, organization, analysis and interpretation of data. As such, these best practice guidelines apply to both the formulation of forensic anthropology methods (basic research) and their applied use in forensic casework. These applied areas primarily include, but are not limited to, stature and age estimation, sex and ancestry classification, and personal identification.

Practitioners should implement these guidelines to the fullest extent as applicable, practical and appropriate. In the absence of specific guidelines or procedures, the principle, spirit and intent should be met.

3.0 General Principles

Forensic anthropologists should:

- Possess a working knowledge of the statistical methods that they apply, including an understanding of the statistical assumptions and limitations.
- Be familiar with the strengths and weaknesses of the data they analyze, whether collected from their own samples or from pre-existing databanks.
- Exercise informed judgment on what statistical methods to use and what weight should be assigned to the results of the statistical procedures employed.
- Be alert to cognitive bias that may affect the recording, analysis and/or interpretation of data and limit its influence by working in the blind whenever pertinent.

4.0 Best Practices

4.1 Overarching and Core Procedures

4.1.1 Sampling

Adequacy of reference samples is crucial to the legitimacy of statistical results. A ‘reference sample’ is used to estimate a population parameter(s) and for this reason is sometimes called the ‘calibration sample.’

Statistical models employed for hypothesis testing are only as good as the reference sample upon which the data are based. In general, samples should be large and randomly drawn from their population (with respect to the traits being measured) to meet the assumptions of most statistical models and parametric statistical significance tests. Where the differences within and/or between groups are of practical consequence, the reference samples should be relevant to the case at hand (e.g., same temporal period, sex, age, and ancestry). The analytical notes should clearly document what reference samples were used (e.g., give citations to the studies employed).

Model performance should ideally be tested using an independent sample (i.e., holdout group). If the reference sample used to derive the estimation model is used for validation, appropriate statistical methods should be employed to minimize bias (e.g., leave-one-out classification).

4.1.2 Data Collection and Organization

Data should be:

- Collected as prescribed by the methods for which the data are to be employed.
- Collected to an acceptable and meaningful level of precision (e.g., appropriate number of significant figures or decimal places).
- Collected using accurate, performance-checked, measuring instruments when metrically established.
- Organized in an orderly manner conducive to technical review / verification.

When appropriate and permissible, anonymized raw data should be submitted to open-access anthropological data repositories to support future research and method improvements.

4.1.3 Statistical Assumptions

Whatever statistical method is employed, the assumptions of the method (e.g., normality, equal variances, random sampling, independence of samples, data type) should be met by the data.

4.1.4 Error and the Uncertainty of Measurement

Any potential problems arising from errors (e.g., intraobserver, interobserver) and/or uncertainty of measurement (e.g., sampling, preservation state) that may affect the accuracy and/or reliability of the test results should be recorded.

Point estimates should be accompanied by clearly labeled prediction intervals. Other estimation and classification results should be accompanied by an indicator of certainty, such as the positive or negative predictive value, correct classification rate, posterior probabilities, and/or typicality probabilities.

4.1.5 Multiple Variables

Multivariate statistics should typically be employed for multivariate data to: maximize the ability to detect differences, explain variation within the data set, mitigate problems of correlation among variables, and reduce the opportunity for Type I error.

When multiple variables are employed to construct prediction or classification models, care should be taken to avoid model over-fitting, especially when employing small samples.

4.1.6 Null-Hypothesis Tests

For research and development of methods, comprehensive exploratory and descriptive data analysis should be conducted prior to null-hypothesis testing to include calculation of moment statistics (mean, variance, kurtosis, and skewness) and/or visualization of the data (i.e., plotting).

The criteria used for setting statistical significance or power should be explicitly stated. Note: universal criteria do not exist for establishing statistical significance or defining acceptable levels of power. As a general rule, p-values should be low and power should be high. Commonly employed p-value thresholds are: 0.05 and 0.01. A power $\geq$0.80 is typically favored. In casework applications, the significance levels recommended in the published test protocol should be documented (whether utilized or not).

The p-values should be reported to an acceptable and meaningful level of precision (i.e., appropriate number of decimal places). P-values should be used as an indicator of the strength of evidence that the data depart from the null-hypothesis. P-values should not be confused with measurements of the strength of an effect (i.e., low p-values only indicate evidence of some effect or that the effect is not nil). Effect magnitudes should be calculated using an appropriate strength of association (e.g., r^2, η^2) and/or effect size statistic.

4.1.7 Classification

Classification includes methods of finding the most similar group or individual in a reference sample. When a classification model is employed in forensic casework, the test result should be accompanied by the model's predictive values, posterior probabilities and/or typicality probabilities. These statistics should be regarded as approximate for the case at hand because all models are estimations.

Classification functions will always indicate the most similar group or individual. Therefore, when the unknown's true group or reference data is not represented in the reference sample(s), analysts should be cognizant that misleading results may be produced.

Typicality probabilities measure the deviation of the data from the null hypothesis that the individual comes from a particular group. Since atypical individuals may be encountered in forensic casework (to include individuals with pathological conditions) care should be exercised when interpreting typicality probabilities. If the typicality probabilities for all groups are low and the reference samples are truly applicable, a thorough check for measurement errors is prudent.

Frequentist statistics should not be confused with Bayesian statistics. Bayesian analysis uses prior probabilities, which should be carefully formulated and explicitly described, justified, and documented.

4.1.8 Casework Selection of Methods

In many circumstances, multiple methods will exist for estimating the same variable. These methods should, therefore, be applied in a prioritized order depending upon their utility (i.e., reliability, applicability and probative value). This principle must, however, be weighed with the unacceptable practices listed in 5.0. Expert opinion should be used to decide how many and which tests are pertinent.

Greatest interpretative weight should be given to estimation models with high correlations and low standard errors.

5.0 Unacceptable Practices

The following practices are considered unacceptable and should be avoided:

- Using statistical procedures that the user does not understand sufficiently to apply and interpret the results.
- Uncritically accepting the results of a single statistical test as a ground truth.
- Employing conventional critical values as a mechanical and dichotomous decision protocol for accepting or rejecting hypotheses.
- Using the p-value as an indicator of strength of an effect or proof of a meaningful difference.
- Blind acceptance of statistical results without regard for their clinical or practical significance.
- Use of multiple univariate tests without concern for increased frequency of Type I errors.
- Use of reference samples with poor applicability to the case at hand (i.e., are not representative).
- Ad hoc formulation of priors when using Bayesian statistics.
- Reporting point estimates without explicit prediction intervals.
- Using the correct classification rate as the primary indicator of method performance without regard for the positive or negative predictive values.
- Application of multiple methods without prioritization for reliability, validity, and probative value.