B978-0-12-815764-0.00003-4, 00003

CHAPTER

3.1 Frequentist approach to data analysis and interpretation in forensic anthropology

Minas Sifakis[a], Michael N. Kalochristianakis[b], Julieta G. García-Donas[c], Oguzhan Ekizoglu[d,e], Elena F. Kranioti[f]

[a]The Alan Turing Institute, London, United Kingdom [b]Medical School, University of Crete, Heraklion, Crete, Greece [c]Centre for Anatomy and Human Identification, School of Science and Engineering, University of Dundee, Dundee, Scotland, United Kingdom [d]Department of Forensic Medicine, Tepecik Training and Research Hospital, Izmir, Turkey [e]Centre Universitaire Romand de Médecine Légale, Lausanne—Genève (CURML), Lausanne, Switzerland [f]Forensic Medicine Unit, Department of Forensic Sciences, Faculty of Medicine, University of Crete, Heraklion, Crete, Greece

Introduction

Forensic anthropology as forensic medicine deals with questions that are likely to end up in court and need to be answered with accuracy. In the United States the implementation of the Daubert standards (Daubert v. Merrell Dow Pharmaceuticals, Inc, 1993) has produced a change in scientific testimony by integrating specific queries such as the method testability; the method scrutiny through peer review and publication, known as potential error rates; and method acceptance by the scientific community. Although the previously accepted Frye Standards (Frye v. Unites States, 1923) for scientific evidence established the necessity of the "general acceptance" of a method, Daubert placed more weight onto the scientific method employed by the experts rather than their qualifications and reputation (Ousley and Hollinger, 2012). According to Daubert, the potential error rates of a method need to be known to evaluate its validity and reliability. Thus forensic anthropologists are encouraged by this legal framework to choose a method appropriate for the specific case.

"Error" from the mathematical or statistical point of view refers to the difference between a computed or measured value to the true value with the deviation from the calculated and correct value being stated by the standard error or prediction intervals (Christensen et al., 2014). For instance, the recommended error for a sex estimation method is less than 5%, that is, the accuracy of the method needs to be over 95% for the classification of a given individual. Nevertheless, the expert must acknowledge,

Statistics and Probability in Forensic Anthropology. https://doi.org/10.1016/B978-0-12-815764-0.00003-4

Obertova, 978-0-12-815764-0

B978-0-12-815764-0.00003-4, 00003

when providing expert witness testimony, that the statistical error is a reflection of the data represented by the sample understudy with its specific variability (Christensen et al., 2014). In practice the selection of a method is subject to other factors such as the preservation of the remains, time, or availability of equipment. For example, if the only method available has a low accuracy, the forensic expert is expected to acknowledge, report, and clearly present the limitations (Christensen and Crowder, 2009).

p0020 The Scientific Working Group for Forensic Anthropology underlines that statistics are intended to "inform scientific inferences via the collection, organization, analysis, and interpretation of data" (SWGATH, 2012). The main questions relate to the estimation of the biological characteristics (age, sex, stature, and ancestry) and to personal identification, but other queries may also emerge, such as the differentiation of peri- and postmortem trauma. The answers are typically provided using inferential statistics, which amounts to any attempt to use properties of a sample to fit an approximate data model and to apply this model to support statements about the general population from which the sample derives. Inferential statistics is divided into two branches: frequentist statistics (which is the subject of this chapter) and Bayesian statistics. Both approaches are based on the study of the properties of past observations to extract information, inferring or predicting characteristics of new data. Yet the mathematical theory, the definition, and general concept of probability differ between frequentist and Bayesian statistics. In the frequentist approach the unknown properties are assumed to be perfectly knowledgeable but unknown to us. If one had access to an infinite number of observations, they could be established without doubt. In the Bayesian approach, the randomness is part of the data-generating process, which is inherently random, and hence, no matter how large the sample, only probabilistic statements can be made about it.

p0025 This paper will showcase the frequentist approach by analyzing an example dataset of morphological features of the tibia in seven Mediterranean populations regarding sex and ancestry estimation. This sample was collected for a previous study (Kranioti et al., 2019), and for the purpose of this example, 100 individuals of Southern European (SE) ancestry and 100 individuals of Turkish (TU) ancestry were randomly selected from the original dataset. Three measurements from the tibia were included in the analysis: maximum length (ML), upper breadth (UB), and lower breadth (LB). The SE group consisted of Greeks, Greek-Cypriots, Italians, Spanish, and Portuguese individuals. Both SE and the TU samples were equally represented by males and females. All calculations were undertaken using the statistic language R and its open source libraries and extensions. Examples of code, written in the R programming language, is available at http://dev.med.uoc.gr/forensic/tbcl/chapter.jsp.

s0020 Exploratory data analysis (EDA)

p0030 The first step in any statistical analysis is to gain basic insight into the data by calculating summary statistics such as minimum, maximum, median, mean values, and quantiles. Summary statistics provide useful information regarding several properties of the data such as the range of each variable, the balance or imbalance of categorical variables, and the symmetry of their distributions. Table 1 presents the

B978-0-12-815764-0.00003-4, 00003

t0010 **Table 1** Summary statistics of the example dataset (measurements in mm).

	ML	UB	LB
Minimum	277.0	49.00	33.51
First quartile	330.0	68.45	44.00
Median	349.5	72.05	47.85
Mean	348.0	72.45	47.56
Third quartile	366.0	76.62	51.00
Maximum	411.0	86.60	59.90

ML, *maximum length;* UB, *upper breadth;* LB, *lower breadth.*

t0015 **Table 2** Correlation matrix of the example dataset.

	Ancestry	Sex	ML	UB	LB
Ancestry	1.000	0.000	0.03	0.129	0.533
Sex	0.000	1.000	−0.511	−0.670	−0.505
ML	0.031	−0.511	1.000	0.575	0.480
UB	0.129	−0.970	0.575	1.000	0.668
LB	0.533	−0.505	0.480	0.668	1.000

ML, *maximum length;* UB, *upper breadth;* LB, *lower breadth.*

summary statistics for the example dataset. Table 2 shows the correlation among the variables by using the correlation coefficient as a measure of the degree of linear association (collinearity) between a pair of variables.

p0035 The next step in data exploration usually involves the visualization of the dataset, such as using histograms or boxplots that reveal the distribution of individual variables and identify potential outliers. Paired scatterplots may be used to identify potential relationships between variables. For the example dataset the research question concerns the estimation of sex and ancestry from the measurements of the tibia; hence, two sets of graphs, for sex and ancestry, respectively, were produced (Fig. 1).

p0040 Fig. 1A reveals sizable differences in the distribution for all variables with respect to sex, and therefore the hypothesis that sex can be inferred from these tibial measurements seems plausible. Fig. 1B presents a less clear-cut situation, where LB measurements seem to differ between groups of different ancestry but UB and ML do not.

p0045 The summary statistics in Table 1 indicate a symmetric distribution of the data; the mean is symmetrically located between 25% and 75% quartile. Some of the histograms in Fig. 1 seem to be slightly skewed, such as in the case of ML measurements for males, although this could be the result of the random sampling process (given the sample size). Most histograms appear to be compatible with the normal distribution assumption. The histograms in Fig. 1 support the hypothesis of equal variance between groups—in less clear-cut cases, this assumption could be verified using

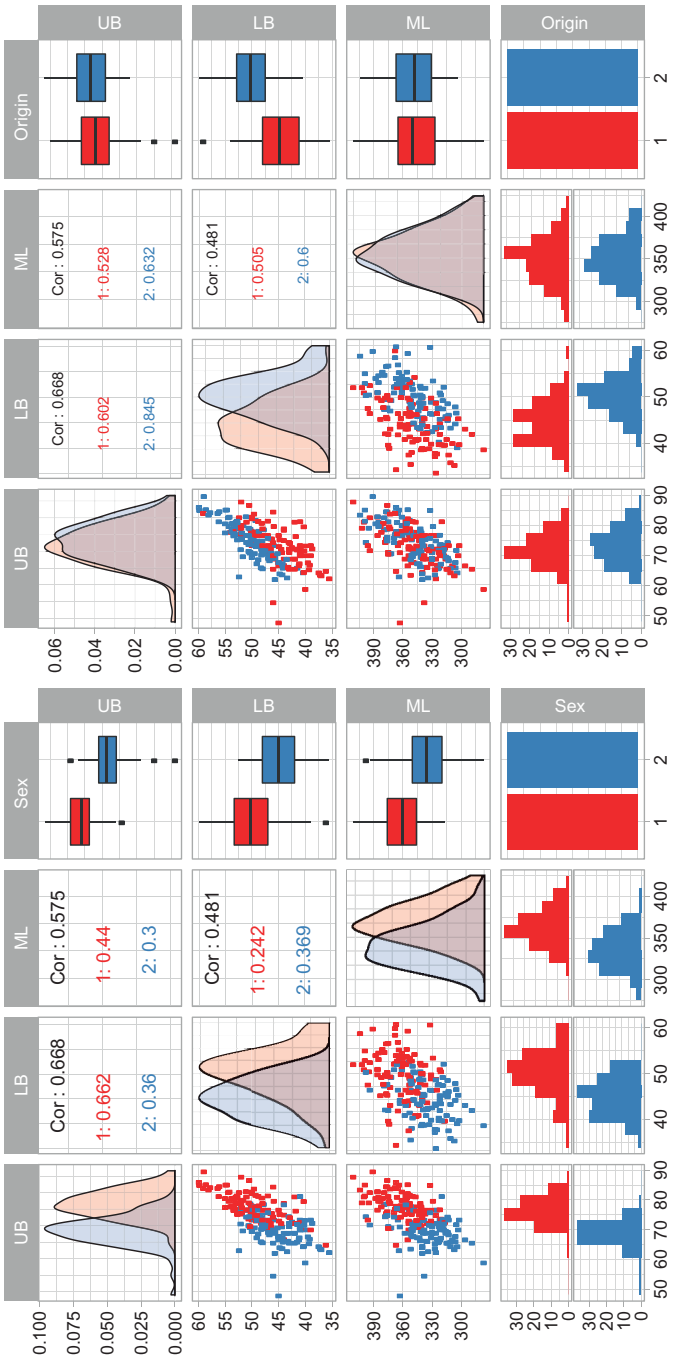

FIG. 1

f0010 Histograms, scatterplots, and boxplots for measurements of the tibia of 200 individuals from seven Mediterranean populations. (A) Females (red) and males (blue) and (B) individuals of South European (red) and Turkish (blue) ancestry.

B978-0-12-815764-0.00003-4, 00003

the F-test. The boxplots for UB in Fig. 1 indicate that there are two isolated measurements at the extreme of the distribution that could be considered as possible outliers or measurement errors. Inspection of the two data points did not reveal any apparent abnormalities, and hence, they were retained, identified as possible high leverage points. The calculated correlation coefficients of all pairs of independent variables were high (Table 2).

p0050 The original dataset, the one from which the example dataset was sampled, was randomly collected from various sites in the Mediterranean Basin and is believed to be a representative, nonbiased sample from the populations involved (Kranioti et al., 2019). Individual specimens were collected from areas and periods with no indications of abnormal population effects, such as epidemics, or natural catastrophes. It is therefore reasonable to assume that the independence assumption holds and that the sample is representative of the population.

s0025 Hypothesis testing

p0055 Hypothesis testing sits at the core of the traditional frequentist approach to inference. It is based on comparisons of properties that characterize data samples and assesses the observed variation by casting it as a decision problem between incompatible hypotheses, the null hypothesis (H0) and the alternative hypothesis (HA). The former often represents "the hypothesis of no difference," while the latter supports the existence of a difference (Curran, 2013). Thus, to explore in the example whether there are metric differences in the tibiae with respect to sex, H0 can be phrased as "There are no differences in the means of tibial measurements between males and females." HA assumes that differences exist. Analogously, when exploring the differences in ancestry of the individuals, the H0 would be "There are no differences in the means of measurements between individuals of South European and Turkish ancestry."

p0060 Having formulated the two hypotheses, the next step is to decide on a statistic, which characterizes a property to be tested. For our example dataset, this could be the difference of the mean value of the maximum length (ML) of the tibia for males and females. H0 is evaluated by calculating the significance of the statistic and by interpreting the result. Significance is often expressed as a *p*-value reflecting "the incompatibility of the data with the null hypothesis." The smaller the *p*-value, the greater the statistical incompatibility of the data with the null hypothesis, assuming that the assumptions used to calculate the null hypothesis hold (Wasserstein and Lazar, 2016).

p0065 The choice of what constitutes a large or a small *p*-value is arbitrary and depends on the acceptable risk of rejecting H0 when it is in fact true. This is called a type I error, and its probability is called significance level of the test, denoted by the Greek letter α. In contrast, accepting H0 when it is in fact false or rejecting HA when it is in fact true comprises a type II error, denoted be the Greek letter β. Setting the value of the significance level is an important decision that examiners make, one that should be made before the actual analysis. In practice, researchers often reject the H0 if the *p*-value is smaller than 0.05 or 0.01, although the results of recently published

B978-0-12-815764-0.00003-4, 00003

metaanalysis studies (Ioannidis, 2019) indicate that thresholds between 0.01 and 0.005 may be more appropriate. Decreasing the significance level reduces type I errors, but it increases the rate of type II errors, since it becomes more difficult to reject H0 even when it is false. This effect is particularly pronounced when working with small- to medium-sized samples. Another related concept is the power of the test, the probability that H0 is rejected when it is indeed false. The power of the test equals 1-β. Decreasing the significance level increases type II errors and reduces the power of the test.

s0030

Comparing two independent samples and the *t*-test

p0070 The most common question arising in forensic anthropology is inferring whether two samples come from the same population and, if so, to quantify the degree of certainty of this statement. This can often be reduced to evaluating whether the distributions of the sample means of the given samples are statistically distinguishable. According to the central limit theorem, for large number of points within each of the groups, the sampling distribution is well approximated by the normal distribution. A *t*-test is then capable of evaluating the distribution of the sample means for the two groups by examining if the difference of their means rests within a predefined confidence interval of the normal curve. The *t*-test statistic, denoted t_d, is the ratio of the difference of the group means to the normal variability across the groups, the standard error (SE).

p0075 For our dataset, let $Y_p = y_{p1},y_{p2},\ldots,y_{pn}$, $X_p = x_{p1},x_{p2},\ldots,x_{pn}$ be the measurements of a quantity p in {UB, LB, ML} for female (Y) and male (X) individuals. Let $\sigma_{px} = \sigma_{py} = \sigma_p$ be their common standard deviation and μ_{px}, μ_{py} the sample means of the measurement for p. The hypotheses of each test can be mathematically formulated as follows:

$$H_0 : \mu_X^p - \mu_Y^p = 0$$

$$H_A : \mu_X^p - \mu_Y^p \neq 0$$

p0080 The model assumptions can also be expressed in the following equations:

$$y_i^p, x_i^p \, i.i.d., X^p \sim \mathrm{N}\left(\mu_X^p, \sigma_X^P\right), Y^p \sim \mathrm{N}\left(\mu_Y^p, \sigma_Y^P\right), \sigma_X^p = \sigma_Y^p = \sigma^p$$

p0085 To calculate the SE, since the variance of the general population ($\sigma 2$) is unknown, one can use a pooled estimate of the variances of the samples:

$$SE = s_{pool}^p \sqrt{\frac{1}{n} + \frac{1}{m}},$$

where

$$s_{pool}^p = \frac{(n-1)s_X^p + (m-1)s_Y^p}{m+n-2}$$

$$s_X^p = \frac{\sum_{i=0}^{n} \left(x_i^p - \mu_X^p\right)^2}{(n-1)}, s_Y^p = \frac{\sum_{i=0}^{m} \left(y^p - \mu_Y^p\right)^2}{(m-1)}$$

B978-0-12-815764-0.00003-4, 00003

p0090 The *t*-test statistic that follows a *t*-distribution with $(df) = n + m - 2$, degrees of freedom is then calculated as follows:

$$t_d^p = \frac{\mu_X^p - \mu_Y^p}{SE}$$

p0095 The resulting *p*-value is determined by comparing the calculated value of the test statistic to the distribution of the values that would occur, and *H*0 should be true:

$$p = Pr(t \geq t_d \vee H0 = \text{True})$$

p0100 If the calculated *p*-value is smaller than the required significance level, α_0, *H*0 is rejected. This test is two sided, since it considers the absolute difference of the means and not whether the difference is positive or negative. When the sign of the difference of means is part of the hypothesis being tested, a one-sided test needs to be used.

p0105 Having calculated the standard error and the degrees of freedom, one can use tabulated values of the cutoff of the *t*-distribution to obtain $t\alpha 0_{df}$ and construct a $(1 - \alpha_0)\%$ confidence interval (CI) for the difference of means. The corresponding formula is

$$(1 - \alpha 0)\% CI = (\mu_X - \mu_Y) \pm t_d^a SE$$

p0110 A confidence interval provides the range of plausible values for point estimates of the data. A 95% confidence interval for the differences of the mean of two groups indicates that, if we took many samples and formed the 95% CI for the mean of each sample, 95% of those intervals would contain the true difference of the means. Confidence intervals are very useful for quickly assessing compatibility of new samples with a wider population or agreements with the requirements of regulations. It is recommended for CIs to be calculated as part of any statistical analysis since they allow assessing the degree of compatibility of a measurement or sample with a source population as opposed to a simple yes/no answer obtained by setting an arbitrary *p*-value threshold.

p0115 For our example dataset, it is necessary to conduct three tests for each H0, one for each tibial measurement (ML, UB, and LB). Tables 3 and 4 summarize the results of the *t*-tests. All tests were two sided, since it makes no difference for our purpose if any potential difference in the means is positive or negative.

t0020 **Table 3** Overview of the *t*-test results for ML, UB, and LB measurements with respect to sex.

***t*-Test**				
	Difference of means	***t*-Metric**	***p*-Value**	***df***
ML	25.79	8.37	1e−14	196.50
LB	7.82	8.23	2e−15	191.62
UB	5.09	12.71	2e−16	196.60

ML, *maximum length;* UB, *upper breadth;* LB, *lower breadth.*

B978-0-12-815764-0.00003-4, 00003

t0025 **Table 4** Overview of the *t*-test results for ML, UB, and LB measurements with respect to ancestry of the individuals.

***t*-Test**				
	Difference of means	***t*-Metric**	***p*-Value**	***df***
ML	1.57	−0.44	0.66	196.76
LB	5.36	−8.87	4e−16	196.47
UB	1.51	−1.83	0.07	197.02

ML, *maximum length;* UB, *upper breadth;* LB, *lower breadth.*

p0120 In Table 3 the calculated *p*-values in all cases are much smaller than the requested significance level of 0.01. At the 95% CI the difference of the means is positive for all cases. More specifically the minimum and maximum difference is 19.7 and 31.9, 6.62 and 9.05, and 3.87 and 6.30 for ML, LB, and UB, respectively (also all positive). The aforementioned numbers can be acquired directly from the output of the t.test() function. It is thus unlikely that the observed differences in the means of the measurements between males and females can be attributed to chance.

p0125 The results of the *t*-tests regarding the ancestry of the individuals indicate that the *p*-value for LB is much smaller than the requested significance level of 0.01, while the *p*-values for ML and UB are not (Table 4). At the 95% CI the minimum and maximum differences of the means are −8.63 and 5.49, 3.12 and 0.12, and −6.56 and −4.18 for ML, UB, and LB, respectively. The difference of means from Table 4 can also be expressed as a ratio to the standard deviation of the data; this yields 0.062, 0.259, and 1.250 for ML, UB, and LB, respectively. These values are very small for ML and UB, and thus only for the case of LB, it is likely that the observed differences in the means of the measurements about ancestry cannot be attributed to chance.

s0035 Deviation from normality

p0130 When the size of the sample population is larger than 30 observations, the t-test is relatively robust to small deviations from normality. For larger sample sizes (100 or more observations), the *t*-test is relatively robust to moderate deviation from normality. When large deviations from normality are suspected, especially when working with small sample sizes, it is recommended to use nonparametric tests. These tests do not require any assumption on the distribution of the data and rely on different techniques to assess the *p*-value from the value of the test outcome, such as ranking. Nonparametric tests are often more robust to outliers. One of the most well-known nonparametric tests is the Mann–Whitney test (Rice, 2008).

s0040 Estimating unknown parameters

p0135 Estimating the value of an unknown quantity, a dependent variable, based on other known quantities, and independent variables or predictors, for the same specimen requires the introduction of a mathematical model. Although there are various

B978-0-12-815764-0.00003-4, 00003

classes of models that can be used for this task (Bishop, 2006), regression models are by far the most widely used and will thus be the focus of this section. Regression models can be divided into two broad categories: those involving continuous dependent variables and those with categorical dependent variables.

Estimating continuous variables: Linear regression

The simplest model that estimates a continuous unknown parameter based on the values of other quantities that are suspected to be relevant is a linear model with zero mean and an uncorrelated error term with constant variance:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

In the equation, y_i is the variable we wish to predict for specimen i, x_{ik} is the kth variable for specimen i, β_0 is the intercept (also the mean value of y in the population), $\beta_{1\ldots k}$ is the regression coefficients for variables $1\ldots k$, and ε_i is an uncorrelated, normally distributed, random variable with zero mean and constant variance. The regression coefficients quantify how much an increase in variable x_{ik} will influence the unknown variable y_i if all the other variables are kept constant. Calculating the parameters of the model is usually carried out using the least squares method (Rice, 2008), which minimizes the sum of the squared residuals (RSS), the differences between the predictions of the model, and the true values of the dependent variable:

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - b_0 + b_1 x_{i1} + \ldots b_k x_{ik})^2$$

If the sample meets the conditions of independence, normality, and homogeneity of variance of the residuals and there is a linear relation between the independent variables and the predictor, we can use the fit statistics to infer the regression coefficients. P-values may be used to assess the significance of individual regression coefficients and to evaluate the stability of the estimates in terms of standard errors, and F-test can be used to check if a subgroup of or all coefficients are zero. To further evaluate the quality of the model fit, one has to inspect the size and distribution of the residuals (they should be randomly distributed around the zero line, and they should have constant variance). We also need to assess the value of the coefficient of determination, R^2, which measures the percentage of variance in the data explained by the model. If some of the assumptions for fitting the model are not satisfied, the fit statistics cannot be relied upon, and therefore additional validation is necessary. It should be emphasized that the use of model validation for evaluating the quality of the fit and the performance of the model is always recommended.

For problems with more complex functional relations among the variables, direct extensions of linear regression involving polynomial and interaction terms, regression splines, or kernel regression can be used (Hastie et al., 2008; Marra and Radice, 2010). If deviation from normality poses an issue, especially in small sample sizes and/or presence of sizeable outliers in the data, one should use more robust regression methods such as median least square regression (MLS) or the MM estimator (Maronna et al., 2006).

B978-0-12-815764-0.00003-4, 00003

s0050

Estimating categorical variables: Logistic regression

p0160 When working with binary dependent variables, the most widely used model is the logistic regression model. Logistic regression estimates the ratio of the probability that an event will occur to the probability that it will not occur. This ratio defines the odds of an event in statistics. The estimation is achieved by fitting a linear model to the logarithm of the odds of the two scenarios. In our example dataset a logistic regression model could be used, for example, to calculate the ratio of the probability that a certain specimen belonged to a female over the probability that it belonged to a male. The logistic regression model equation is

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

where π_i is the probability of the specimen i belonging to the first class and $(1-\pi_i)$ is the probability of the specimen i belonging to the second class. The rest of the equation is identical to the one for linear regression. Rearranging the earlier equation, one obtains the equation for predicting the probability of a class:

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}}}$$

p0165 The class prediction is obtained by setting a threshold, π_{th}, and assigning all specimens with $\pi_i > \pi_{\text{th}}$ to the first class and all remaining to the second. The most commonly used threshold value is $\pi_{\text{th}} = 0.5$. However, particularly when there is asymmetry between the different types of errors, a higher or lower threshold may yield more appropriate results. In that case the appropriate value of the threshold is determined empirically, by evaluating model performance on a range of thresholds using k-fold cross validation (see next section).

p0170 In the case of multilevel dependent variables, for example, when classifying a specimen to one of j possible groups, one can fit j-1 different one-versus-all binary models and select the class with the largest pairwise probability. Alternatively, one can use multinomial regression, often called *softmax* regression, to calculate the probability of each class under the model (Hastie et al., 2008).

p0175 Logistic regression belongs to a class of models known as generalized linear models (GLMs) (Dunteman and Ho, 2006). There is no analytic solution for the estimation of the parameters of GLMs; thus calculations are carried out with the aid of software, using methods such as the iterative reweighted least squares. Logistic regression assumes that the observations are independent, that the sample is representative of the wider populations and free of numerous outliers, and that the log odds are linearly related to the predictors (Hastie et al., 2008). When these conditions are satisfied, inference on individual regression coefficients is possible using the calculated p-values. Hypothesis testing, for the model as a whole, can be carried out using the Hosmer-Lemeshow test. The quality of the fit is assessed using the deviance D of the model, which is equivalent to the sum of squared errors. Smaller values of the deviance correspond to a better fit. Although deviance can be used to compare models, there is no straightforward interpretation of deviance values in isolation

B978-0-12-815764-0.00003-4, 00003

(contrary to R^2). Compared with linear regression, residual plots are less informative in logistic regression. However, the plots of the deviance residuals and the Cook's distances (Seber and Lee, 2003) should be inspected to identify influential observations and/or potential outliers. A large value of Cook's distance indicates an influential observation, so observations with a large Cook's distance should be inspected to determine if they correspond to outliers (in which case they should be removed from further analysis) or if they simply represent the extremes of the data distribution.

s0055

Model selection

p0180 Although in principle one could fit a model using all available independent variables, in practice, it is desirable to build a model that contains only the informative variables, those crucial to the estimation of the dependent variable. Exploring such models and selecting the most appropriate per case is a valuable tool for forensic anthropologists and forensic scientists to gain insight into the possible data generating processes and to provide generalization for unknown cases.

p0185 For the tibia dataset, since both dependent variables are binary, binary logistic regression is a natural model selection. Selecting the variables that will be included in the model is in this case straightforward, but in more complex situations where the number of variables can be large and their importance not evident, the model selection procedure can be complex or may require expertise and experience that is not readily available. In such cases, automatic model selection methods, such as the stepwise selection methods and regularized regression (LASSO, Ridge regression (see Hastie et al., 2008)), can be applied.

p0190 The stepwise selection method, which will be presented as an example here, works by iteratively adding and removing predictors, each time refitting the model and evaluating its performance on the dataset until the optimal combination of variables is established. The evaluation of the model's performance may rely on metrics such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) (Hastie et al., 2008). Both of these metrics balance the incremental performance gain of adding one more variable to the model with a complexity penalty. The smallest values of the two metrics correspond to the preferred models.

p0195 The output of a logistic regression fitted on the tibia dataset produces two outcomes: the class predictions, that is, male/female or SE/Turkish, and the probability of the specimen belonging to each of the two classes. The actual model equations are the following:

$$\log\left(\frac{\pi_{\text{Female}}}{1-\pi_{\text{Female}}}\right) = 36.1274 - 0.0216*\text{ML} - 0.3948*\text{UB}$$

$$\log\left(\frac{\pi_{\text{SE}}}{1-\pi_{\text{SE}}}\right) = -5 - 0.024*\text{ML} + 0.559*\text{LB} - 0.184*\text{UB}$$

B978-0-12-815764-0.00003-4, 00003

t0030 **Table 5** Logistic regression results of best fit for the sex estimation model, AIC = 156.6.

Coefficients				
	Estimate Std.	**Error *z***	**Value**	***Pr*(>|*z*|)**
(Intercept)	36.127	4.989	7.241	4e − 13
UB	−0.394	0.064	−6.132	9e − 10
ML	−0.021	0.010	−2.100	0.035

t0035 **Table 6** Logistic regression results of best fit for the ancestry estimation model, AIC = 184.41.

Coefficients				
	Estimate std.	**Error *z***	**Value**	***Pr*(>|*z*|)**
(Intercept)	4.997	2.752	−1.816	0.069
UB	−0.184	0.0564	−3.259	0.001
ML	0.024	0.009	−2.534	0.011
LB	0.559	0.080	6.959	3e − 12

p0200 The results of the best fits for sex and ancestry estimation models, respectively, are shown in Tables 5 and 6. The tables summarize the output of the GLM function for R. The actual code is available in http://dev.med.uoc.gr/forensic/tbcl/chapter.jsp, examples 3 and 4.

p0205 The results of the model selection exercise may seem counterintuitive at first. In the sex estimation example, where H0 was rejected for all independent variables, the best identified model does not include all variables since LB was excluded. On the contrary, in the ancestry estimation model, where H0 was rejected for LB, the best model identified includes all three independent variables. The key to understanding how this can be possible lies in the observation that the model assumption of uncorrelated independent variables is violated. The individual t-tests evaluated H0 for each variable in isolation. However, the existence of collinearity between independent variables may imply "information sharing" and possibly redundancy. By evaluating all possible combinations of variables during the model selection, we established that there is redundancy between the variables, and hence, LB was excluded from the model. The inclusion of all variables in the ancestry estimation model is an artifact of ignoring the correlations between the independent variables. Variables that have little significance when evaluated in isolation can contribute to discriminating information in the context of other variables, which is an effect called confounding.

B978-0-12-815764-0.00003-4, 00003

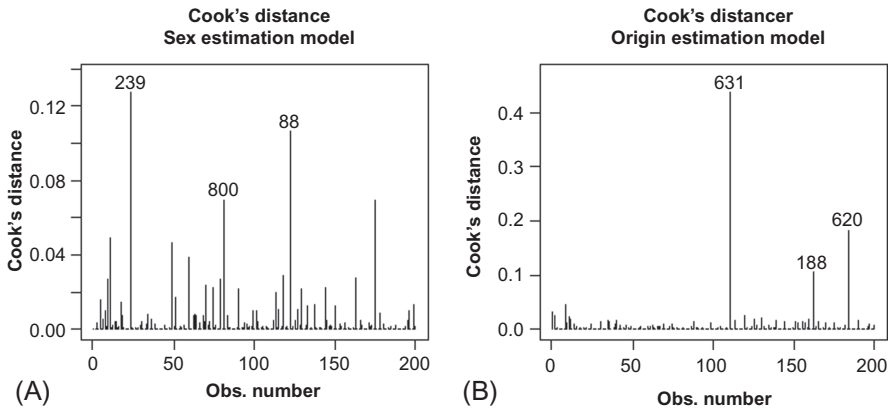

FIG. 2

f0015 Cook's distance plots for (A) sex estimation model and (B) ancestry estimation model.

p0210 As mentioned earlier, evaluating the quality of fit for GLMs is not trivial. Cook's distance plot for the sex estimation model (Fig. 2A) shows two observations (88 and 239), which could be potential outliers. The equivalent plot for the ancestry estimation model (Fig. 2B) shows three such observations (188, 620, and 631).

s0060 Assessing performance: Model validation

p0215 The regression parameters provide no direct information regarding the performance of the model on the actual predictive task. Assessing the model's predictive performance and generalization capacity involves selecting an appropriate performance metric and evaluating this metric on a large, independent sample from the wider population. In practice, however, the requirement for a second independent dataset is often unrealistic. In such cases the sample population could be divided into two, a reference population to be used in the development of the model and a validation population, often called the holdout set. Although this approach provides an independent verification of the model's performance, it does not provide any information about its reliability or about its performance on a different input. A better, although computationally more demanding, approach is to evaluate the model using an iterative resampling method such as k-fold cross validation (James et al., 2015). First the sample dataset is randomly shuffled and split into k parts. One set is set aside to be used for validation, and the model is fit on the other k-1 parts. Once fitted the model is evaluated on the dataset that was not used for the fit. This process is repeated k times, each time using a different subsample as the validation set and the remaining data for model fit, until all combinations are exhausted. This procedure yields k different estimates of the generalization ability of the model that can be used to calculate the mean and standard deviation of the estimates. For $k = 2$, it yields the basic train-validation split described earlier, while for $k = n$, where n is the size

B978-0-12-815764-0.00003-4, 00003

of the dataset, one obtains the leave-one-out cross-validation (LOOCV) method (James et al., 2015). The selection of the performance metric depends on the type of problem, the relevant significance and cost of type I and type II errors, and on the nature of the dataset. Some of the most commonly used metrics include the classification accuracy, the $F1$ score, or the area under the curve (AUC) (James et al., 2015). It is also important for any model evaluation to compare the model's performance with a realistic baseline.

p0220 Before assessing the performance of the models for our example dataset, we investigated whether removing the three observations (188, 620, and 631) with large values of Cook's distance from the training set would have a positive effect in the ancestry estimation model. There was no notable difference in performance; thus the results presented here include all observations.

p0225 The predictive performance of both models was estimated using 10-fold cross validation. The class selection threshold was set to $\pi_{th} = 0.5$. The accuracy of the sex estimation model is 82.5%, with a standard deviation of 8.6%. The accuracy of the ancestry estimation model is 78%, with a standard deviation of 8.9%.

s0065 Discussion

p0230 Frequentist statistics have been applied on several occasions in forensic anthropology. Prediction models regarding sex, ancestry, age, and stature estimation have been developed based on hypothesis testing (e.g., Buckberry and Chamberlain, 2002; Krishan and Sharma, 2007; Walker, 2008; Kranioti et al., 2018). Yet, it is not always clear whether the results are well understood or well interpreted. Let's consider as an example the two questions explored in this chapter. We used metric variables to predict sex and ancestry, both binary variables. The H0 for both outcomes was rejected because variables were found to be statistically significantly different between the groups, and the models developed achieved a cross-validation accuracy of 82.5% for sex and 78% for ancestry. In the first case the tibial measurements were used to infer sex. The specimen can be either male or female. Thus the decision is purely binary; the probability of an unknown tibia that belongs to one of the two groups sums up to 100%. So in that case the result would indicate that the probability that the bone belongs to a female is 95% (vs 5% to a male), that is, it is 19 times more likely that the bone belongs to a female compared with a male. In the case of ancestry though, even if the model was created to provide a binary response (South European or Turkish), there is always the possibility of the person belonging to a group not represented in the dataset (e.g., African). Thus the interpretation and reporting are not as straightforward as for sex estimation. For an answer that is similar as in the sex estimation example, that is, if the result indicates that the probability of the unknown tibia belonging to a Turkish individual is 95% versus 5% to a South European individual, the individual is 19 times more likely to be Turkish than Southern European. However, given the current dataset, no information can be provided regarding other population groups, if the question would be what the chances are

B978-0-12-815764-0.00003-4, 00003

of that specific individual descending from African or Asian populations. Thus the rejection of H0 does not guarantee a solid answer to the ancestry problem. This example demonstrates that even a well-executed statistical analysis does not always lead to efficient problem solving.

p0235 Another important point to note is that frequentist hypothesis testing introduces an asymmetry between the two hypotheses. All calculations and statements are made with respect to the validity of H0. No calculations or statements are made regarding the alternative hypothesis. This is a major difference in comparison with Bayesian statistics where both H0 and HA are tested individually to express the odds. Furthermore, one should understand that in the frequentist approach, failing to reject the H0 does not mean that the expert believes that it is true; it just indicates that the evidence collected may have not been sufficient to establish beyond every reasonable doubt that it is not false. To convey that fact, it is common practice to use double negatives (e.g., "failed to reject the H0") when communicating the results of the frequentist hypothesis testing.

p0240 Due to the aforementioned limitations of hypothesis testing and a growing awareness of its widespread abuse, a transition toward a post-*p*-value statistical approach to frequentist inference is currently underway. Advocates of this approach propose using confidence intervals, effect sizes, and the notion of degree of compatibility rather than the binary acceptance/rejection of hypothesis based on comparison with arbitrary set significance values (Amrhein et al. 2019; Wasserstein et al. 2019).

p0245 In conclusion, taking into account the critical comments, statistical approaches in forensic anthropology and forensic sciences in general should follow a simple rule: the expert should choose the appropriate method according to the available data and questions asked in the given case and ensure to understand and interpret the results supporting their assessment with statistical evidence.

sp0075 *Probability is common sense reduced to calculation.*

Laplace

References

Amrhein, V., Greenland, S., McShane, B., 2019. Scientists rise up against statistical significance. Nature 567, 305–307.

Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer-Verlag Berlin, Heidelberg.

Buckberry, J.L., Chamberlain, A.T., 2002. Age estimation from the auricular surface of the ilium: a revised method. Am. J. Phys. Anthropol. 119, 231–239.

Christensen, A.M., Crowder, C.M., 2009. Evidentiary standards for forensic anthropology. J. Forensic Sci. 54, 1211–1216.

Christensen, A.M., Crowder, C.M., Ousley, S.D., Houck, M.M., 2014. Error and its meaning in forensic science. J. Forensic Sci. 59, 123–126.

Curran, J.M., 2013. The frequentist approach to forensic evidence interpretation. In: Siegel, J., Saukko, P. (Eds.), Encyclopedia of Forensic Sciences, second ed. Elsevier, UK, pp. 286–291.

B978-0-12-815764-0.00003-4, 00003

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 US 579, 1993.

Dunteman, G.H., Ho, M.-H.R., 2006. An Introduction to Generalized Linear Models. SAGE Publications.

Frye v. Unites States, 1923. 54 App. D.C. 46, 293 F. 1013.

Hastie, T., Tibshirani, R., Friedman, J., 2008. The Elements of Statistical Learning. Springer, New York.

Ioannidis, J., 2019. What have we (not) learnt from millions of scientific papers with P values? Am. Stat. 73 (suppl. 1), 20–25.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2015. An Introduction to Statistical Learning With Applications in R. Springer, New York.

Kranioti, E.F., García-Donas, J.G., Can, I.O., Ekizoglu, O., 2018. Ancestry estimation of three Mediterranean populations based on cranial metrics. Forensic Sci. Int. 286, 265.e1–265.e8.

Kranioti, E.F., Garcia-Donas, J.G., Karell, M.A., Cravo, L., Ekizoglu, O., Apostol, M., Cunha, E., 2019. Metric variation of the tibia in the Mediterranean : implications in forensic identification. Forensic Sci. Int. 299, 223–228.

Krishan, K., Sharma, A., 2007. Estimation of stature from dimensions of hands and feet in a North Indian population. J. Forensic Leg. Med. 14, 327–332.

Maronna, R., Martin, R., Yohai, V., 2006. Robust Statistics: Theory and Methods. Wiley Series in Probability and Statistics, England.

Marra, G., Radice, R., 2010. Penalised reression splines: theory and application to medical research. Stat. Methods Med. Res. 19, 107–125.

Ousley, S., Hollinger, R., 2012. The pervasiveness of Daubert. In: Dirkmaat, D.C. (Ed.), A Companion to Forensic Anthropology, first ed. Blackwell Publishing, pp. 654–665.

Rice, J.A., 2008. Mathematical Statistics and Data Analysis. Duxbury Advanced Series.

Scientific Working Group for Forensic Anthropology (SWGATH), 2012. Statistical Methods. Published Documents, Issue date: 08/01/2012.

Seber, G.A.F., Lee, A.J., 2003. Linear Regression Analysis, first ed. John Wiley & Sons, New Jersey.

Walker, P.L., 2008. Sexing skulls using discriminant function analysis of visually assessed traits. Am. J. Phys. Anthropol. 136, 39–50.

Wasserstein, R., Lazar, N., 2016. The ASA statement on p-Values: context, process, and purpose. Am. Stat. 70 (2), 129–133.

Wasserstein, P., Schirm, A., Lazar, N., 2019. The American Statistician. Special Issue: Statistical Inference in the 21st Century: A World Beyond $p < 0.05$, pp. 1–19.

B978-0-12-815764-0.00003-4, 00003

Non-Print Items

Abstract

Forensic anthropologists are requested to give authorities estimates on the biological characteristics of unidentified decomposed remains in an effort to create a physical description that can be compared with a missing person's profile, eventually leading to positive identification. To answer these questions, scientists traditionally follow a variety of statistical approaches such as frequentist and Bayesian statistics in both analysis and interpretation. The main subject of this chapter is to summarize the frequentist statistical approach and to illustrate with examples the rationale in the method selection and interpretation.

Keywords: Forensic anthropology, Forensic medicine, Frequentist statistics, Frequentist inference, Scientific computing