Bayes factor: A useful tool to quantitatively evaluate and compare performance of multiple stature estimation equations

Yangseung Jeong[a],[*],[1], Ashlin P. Harris[b],[1], Omar Ali[c], Yochun Jung[d],[*]

[a] *Department of Biology, Middle Tennessee State University, Murfreesboro, TN 37132, USA*
[b] *Computational Science, Middle Tennessee State University, Murfreesboro, TN 37132, USA*
[c] *Department of Chemistry, Middle Tennessee State University, Murfreesboro, TN 37132, USA*
[d] *Department of Thoracic and Cardiovascular Surgery, Chonnam National University Hospital, Chonnam National University School of Medicine, Gwangju, 61469, South Korea*

A B S T R A C T

When stature estimation of incomplete skeletal remains is necessary, researchers select an estimation equation which will produce the most accurate estimates. The purpose of this study is to propose that, given prior information of a target sample, the Bayes factor can be a useful tool to quantitatively evaluate and compare performance of multiple equations in this regard. This study also explores the best-performing equations to reconstruct statures of Korean War casualties with a demonstration of equation comparisons by the Bayes factor.

Thirty-three sets of stature estimates were generated using different equations based on the osteometric data of the Korean War casualties. The distribution of each set was compared to that of the population (i.e., Korean servicemen during the Korean War) using the Bayes factors and posterior probabilities generated by the R codes in the LearnBayes package.

A higher Bayes factor indicates a closer similarity between the two distributions under comparison. The equation with the highest Bayes factor in this study was Choi et al.'s (1997) humerus equation (bf = 9.84), followed by the femur equation of the same authors (bf = 5.3).

The Bayesian approach has advantages over the traditional frequentist approach primarily based on the *p*-value. Particularly, the Bayes factor can provide practical interpretations on the models under comparison, which allows for a quantitative prioritization of different models. Researchers can obtain more accurate stature estimates of a target sample by using the equation of the highest Bayes factor.

© 2020 Elsevier B.V. All rights reserved.

1. Introduction

As one of the biological parameters representing size of individuals, stature has played an important role in various fields such as bioarchaeology, paleontology, and forensic anthropology [1–17]. In reality, whether living or dead, direct measurement of stature from an individual is not always possible due to his or her incomplete condition antermortem or perimortem (e.g., amputation of lower limbs) as well as due to postmortem taphonomic changes occurring to the individual [18,19]. Thus, extensive effort to reconstruct stature from partial body parts and/or skeletal remains has led to many techniques of stature estimation, which

are generally grouped into one of the two categories: the anatomical method and the mathematical method [20–24].

In order to estimate stature of skeletal remains, it is important to decide which method should be used. In general, regardless of the ancestry or sex of a target sample, the anatomical method, particularly the revised Fully technique by Raxter and colleagues [23], is preferred due to its high accuracy of estimates [24–26]. The anatomical method requires the lengths/heights of the skeletal elements constituting a standing height (i.e., cranial height, body heights of the C2 through L5, height of the first sacral vertebra, femoral physiological length, tibial length, and talus-calcaneus length). Since the skeletal portion of stature is already taken into account, only the soft tissue portion needs to be 'estimated' in the anatomical method, which allows for a smaller prediction error than other types of methods [24,27,28]. However, its application is limited to complete or nearly complete skeletons with all the required elements present [23–28].

* Corresponding authors.
E-mail addresses: yangseung.jeong@mtsu.edu (Y. Jeong), yochuni@naver.com (Y. Jung).
[1] These authors have equally contributed to this manuscript as first authors.

Table 1
Equations available to reconstruct statures of Korean War casualties (unit: cm).

Study	Bones[*]					
	Humerus (n = 128)	Radius (n = 49)	Ulna (n = 34)	Femur (n = 397)	Tibia (n = 239)	Fibula (n = 23)
Choi et al. [42]	$S=4.3H_R+33.32$	$S=3.89R_R+74.12$	$S=3.74U_R+70.78$	$S=2.93Fem_R+36.88$	$S=2.54T_R+73.38$	$S=2.55Fib_R+74.49$
Lee et al. [43]	–	–	–	$S=2.883Fem_R+42.7728$	$S=2.618T_R+75.8742$	$S=2.43Fib_R+83.753$
Lee et al. [44]	$S=5.061H_R+13.3449$	$S=6.829R_R+11.1025$	$S=6.285U_R+11.3027$	–	–	–
Jeong and Jantz [45]	$S=2.514H+85.23$	$S=2.631R+101.243$	–	$S=2.167Fem+69.544$	$S=2.321T+81.488$	–
Lee et al. [46]	–	–	–	$S=2.61Fem_L+54.081$; $S=2.593Fem_R+55.17$	–	–
Pearson [47][**]	$S=2.894H_R+70.641$	$S=3.271R_R+85.925$	–	$S=1.88Fem+81.306$	$S=2.376T_R+78.664$	–
Trotter and Gleser [48]	$S=2.68H_L+83.27$; $S=2.69H_R+82.8$	$S=3.51R_L+83.4$; $S=3.58R_R+80.71$	$S=3.46U_L+78.84$; $S=3.5U_R+76.07$	$S=2.18Fem_L+71.11$; $S=2.12Fem_R+74.03$	$S=2.36T_L+82.54$; $S=2.42T_R+80.36$	$S=2.4Fib_L+80.38$; $S=2.39Fib_R+81.1$
Fujii [49]	$S=2.83H_L+72.908$; $S=2.79H_R+73.242$	$S=3.3R_L+83.401$; $S=3.23R_R+84.296$	$S=3.25U_L+79.201$; $S=3.09U_R+82.587$	$S=2.5Fem_L+53.56$; $S=2.47Fem_R+54.901$	$S=2.36T_L+77.542$; $S=2.47T_R+73.999$	$S=2.55Fib_L+72.97$; $S=2.6Fib_R+70.925$

[*] S – stature; H – humerus; R – radius; U – ulna; Fem – femur; T – tibia; Fib – fibula; subscripts ($_L$ or $_R$) indicate the side of the bones.
[**] When left bones were used, correction factors (+0.42 cm, +0.28 cm, -0.04 cm, and +0.18 cm for the humerus, radius, femur, and tibia, respectively) were added following Pearson [47; p.197].

In case of incomplete skeletal remains, stature is generally estimated by entering one or some bone dimensions into a regression equation, which is referred to as the mathematical method. The equations are devised based on a high correlation between stature and bone dimensions of a reference sample [23,24]. Thus, given the same geographical, temporal, and biological backgrounds of a reference sample as a target sample, the equation is expected to yield an estimate of decent accuracy [27–30].

Then what if there are no equations devised from a reference sample of the same background as a target sample? In this situation, researchers often choose one from currently available equations despite possible discrepancies between the reference and target samples [e.g., 31,32]. On the other hand, if there are multiple equations for which reference samples have the same or similar background as the target sample, researchers would choose one presumably best equation to use. In either situation, it is crucial that the performance of available equations be assessed and prioritized by an objective standard before the one to use is selected. The primary purpose of this study is to propose that the Bayes factor can be used in this regard particularly in cases where prior information of a target sample is known. In order to demonstrate a detailed process of comparing different equations using the Bayes factor, Korean anthropometric and osteometric data will be utilized. Hence, it is the secondary purpose of this study to determine the best-performing equation for this specific Korean sample.

1.1. Bayes factor

In classical or frequentist statistics, the probability of an event is its relative frequency over sufficiently many repetitions. In contrast, under the paradigm of Bayesian statistics, probabilities are reinterpreted as a degree of belief for an outcome. This allows other types of unknowns to be described as random variables, meaning that Bayesian statistics can be used to address topics that might lack a clear definition in the classical case. For instance, a likelihood can be assigned to a scientific hypothesis. Such likelihoods are initially based on a subjective assessment of the probability (the prior probability) and are updated based on further observations, and they can be directly compared to facilitate model selection. The Bayes factor, which depends on the Bayesian definition of the posterior probability for a model, is a ratio of marginal likelihoods for two hypotheses/models and indicates the relative strength of evidence for the two hypotheses/models [33,34]. To select among several models in the Bayesian

context, it is valid to calculate one Bayes factor for each and to choose the model with the highest Bayes factor or posterior probability [35]. Calculating the Bayes factor precisely can be prohibitive due to computational costs or difficulty in setting prior probabilities, in which case the Bayes factor can be estimated with alternate criteria.

2. Materials and methods

To compare performance of multiple equations for a certain target sample, the target sample ideally should consist of individuals from which both reliable stature information and osteometric measurement data are obtainable. In this ideal
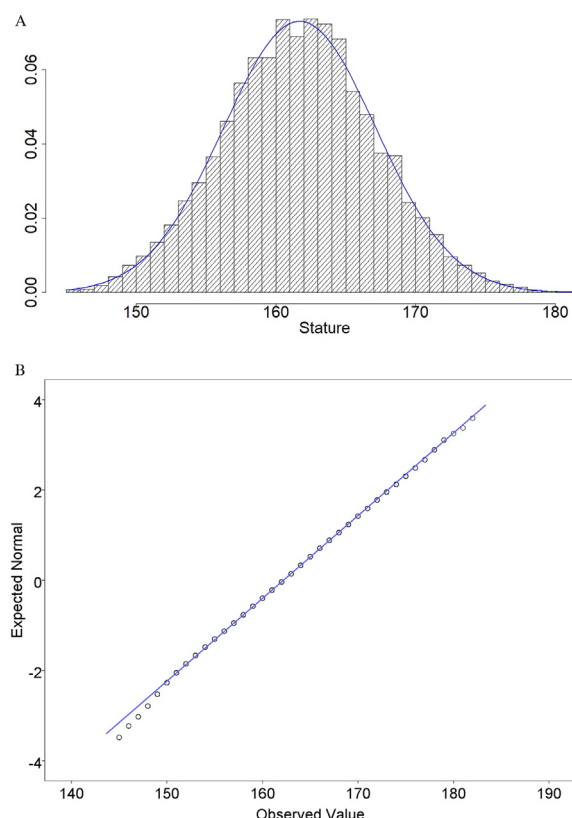

Fig. 1. Histogram (a) and Q-Q plot (b) exhibiting the normal distribution of statures of Korean War servicemen in Park et al. [36].

Table 2
Descriptive statistics of the long bone maximum lengths of the Korean War casualties.

Bone	Side	Mean (cm)	SD (cm)	Min–Max (cm)	p^*
Humerus	Left (n = 76)	30.0	1.4	27.2–33.5	0.542
	Right (n = 73)	30.3	1.3	27.4–34.0	(df = 20)
Radius	Left (n = 31)	22.4	1.3	19.7–25.4	–
	Right (n = 24)	22.5	1.6	19.0–25.3	
Ulna	Left (n = 19)	23.9	1.9	21.2–28.6	–
	Right (n = 16)	24.5	1.3	21.4–26.6	
Femur	Left (n = 241)	42.8	2.2	35.0–49.5	0.511
	Right (n = 241)	42.8	2.0	37.5–48.5	(df = 84)
Tibia	Left (n = 141)	34.3	1.9	28.9–38.4	0.442
	Right (n = 143)	34.3	2.0	28.0–39.6	(df = 44)
Fibula	Left (n = 17)	32.9	1.5	30.0–35.5	–
	Right (n = 11)	33.4	2.7	30.2–38.3	

* Paired t-test was not conducted for the radius, ulna, and fibular due to a small sample size.

condition, multiple sets of stature estimates can be generated using different equations of interest, for which distributions can be compared to the known distribution of the target sample. In fact, some modern skeletal collections, mostly established based on medical school cadavers and body donations (e.g., The Robert J. Terry Anatomical Collection of the National Museum of Natural History of the Smithsonian Institution, William Bass Donated Skeletal Collection at the University of Tennessee), allow for this type of research. However, such skeletal collections with known antemortem information are not available in most countries including South Korea. In this situation, the second-best option will be to compare the distributions of a target sample's estimated statures to the known parameters of the believed-to-be population. In this study, the stature estimates based on the osteometric data from the Korean War casualties and measured statures of the Korean War servicemen were used as examples of the former and latter, respectively. The rationale is that, although the two samples did not consist of the identical individuals, the measured statures and estimated statures are supposed to generate nearly the same distributions due to their identical temporal, geographical, and biological backgrounds (i.e., young Korean males who were born in the early 20th century and died during the Korean War).

2.1. Antemortem stature information of the Korean War servicemen

From December 1952 to March 1953, which was during the Korean War, Park and colleagues [36] collected anthropometric data from 12,081 males between 18–30 years of age (mean age = 23.8 years; SD = 3.1 years), approximately 88% of which were army conscripts and the rest of which were university applicants. Since Park et al. [36] pooled the two groups without an explicit indication of discrepancies and also the number of army conscripts (n = 10,587) considerably exceeded that of university applicants (n = 1,494), on the whole, the reported statistics (mean stature = 162.8 cm; SD = 5.46 cm) were regarded as the population parameters of the Korean War servicemen in this study.

Table 3
Descriptive statistics of the stature estimates by different equations†.

Bone used	Statistics	Choi et al. [42]	Lee et al. [43]	Lee et al. [44]	Jeong and Jantz [45]	Lee et al. [46]	Pearson [47]	Trotter and Gleser [48]	Fujii [49]
Humerus	Mean	163.0		166.0**	161.0**		158.4**	164.0**	157.7**
(n = 128)	SD	5.93		6.98	3.45		3.98	3.70	3.86
	Min–Max	151.1–179.5		152.0–185.4	154.0–170.7		149.9–169.0	156.5–174.3	149.7–168.1
Radius	Mean	161.5		164.5	160.4**		159.9**	161.7	157.2**
(n = 49)	SD	5.42		9.52	3.65		4.58	4.99	4.55
	Min–Max	148.0–172.5		140.6–183.9	151.2–167.9		148.1–168.7	148.7–171.3	145.7–166.0
Ulna	Mean	161.0		162.9				161.5	157.4**
(n = 34)	SD	5.99		10.1				5.43	5.08
	Min–Max	150.1–177.7		144.5–191.1				151.0–177.8	148.1–172.2
Femur	Mean	162.3	166.2**		162.3*	166.0**	161.8**	164.7**	160.6**
(n = 397)	SD	6.13	6.03		4.53	5.44	3.93	4.49	5.20
	Min–Max	139.4–181.9	143.7–185.5		145.4–176.8	145.4–183.3	147.0–174.3	147.4–179.0	141.1–177.3
Tibia	Mean	160.7**	165.9**		161.3**		160.5**	163.6**	158.8**
(n = 239)	SD	4.78	4.93		4.37		4.49	4.52	4.57
	Min–Max	144.5–174.0	149.2–179.6		146.5–173.4		145.2–172.8	148.1–176.2	143.2–171.8
Fibula	Mean	159.8*	165.1**					157.0**	158.1**
(n = 23)	SD	5.20	4.95					6.24	5.29
	Min–Max	151.5–172.2	157.1–176.8					149.1–172.6	149.5–170.5

† One sample t-test indicates a significant discrepancy between the mean of the estimates and the population mean (162.8 cm) at the level of α = 0.05 (*) or α = 0.01 (**).

Table 4
Kolmogorov-Smirnov test results for stature estimates by different equations.

Bone used	Statistic	Choi et al. [42]	Lee et al. [43]	Lee et al. [44]	Jeong and Jantz [45]	Lee et al. [46]	Pearson [47]	Trotter and Gleser [48]	Fujii [49]
Humerus	D	0.681		0.681	0.731		0.666	0.739	0.642
(n = 128)	p-value	0.743		0.743	0.659		0.766	0.646	0.804
Radius	D	0.487		0.487	0.494		0.369	0.364	0.395
(n = 49)	p-value	0.972		0.972	0.968		0.999	0.999	0.998
Ulna	D	0.734		0.734				0.492	0.646
(n = 34)	p-value	0.655		0.655				0.969	0.798
Femur	D	1.165	1.165		0.068	1.003	1.114	1.009	1.157
(n = 397)	p-value	0.132	0.132		0.204	0.267	0.167	0.260	0.138
Tibia	D	0.615	0.615		0.668		0.578	0.573	0.584
(n = 239)	p-value	0.844	0.844		0.763		0.892	0.897	0.885
Fibula	D	0.624	0.624					0.959	0.563
(n = 23)	p-value	0.831	0.831					0.317	0.909

Table 5
Bayes factors and posterior probabilities of each equation.

Bone used	Statistic[*]	Choi et al. [42]	Lee et al. [43]	Lee et al. [44]	Jeong and Jantz [45]	Lee et al. [46]	Pearson [47]	Trotter and Gleser [48]	Fujii [49]
Humerus	bf	9.84[†]		2.2e-05	8.6e-07		7.0e-33	2.3e-02	4.7e-47
	post	0.91		2.2e-05	8.6e-07		7.0e-33	2.2e-02	4.7e-47
Radius	bf	1.77		2.0	2.0e-04		4.1e-04	2.35[†]	9.4e-16
	post	0.64		0.67	2.0e-04		4.1e-04	0.70	9.4e-16
Ulna	bf	1.23		0.91				2.19[†]	3.2e-08
	post	0.55		0.49				0.69	3.2e-08
Femur	bf	5.3[†]	8.3e-27		2.7	6.4e-30	3.0e-05	5.3e-14	1.9e-14
	post	0.84	8.3e-27		0.73	6.4e-30	3.0e-05	5.3e-14	1.9e-14
Tibia	bf	1.7e-09	1.4e-19		1.1e-05		7.7e-13	0.47[†]	6.2e-39
	post	1.7e-09	1.4e-19		1.1e-05		7.7e-13	0.32	6.2e-39
Fibula	bf	0.13	0.53[†]					3.7e-04	8.7e-04
	post	0.12	0.35					3.7e-04	8.7e-04

[*] bf – Bayes factor in support of the null model; post – posterior probability of the null model.
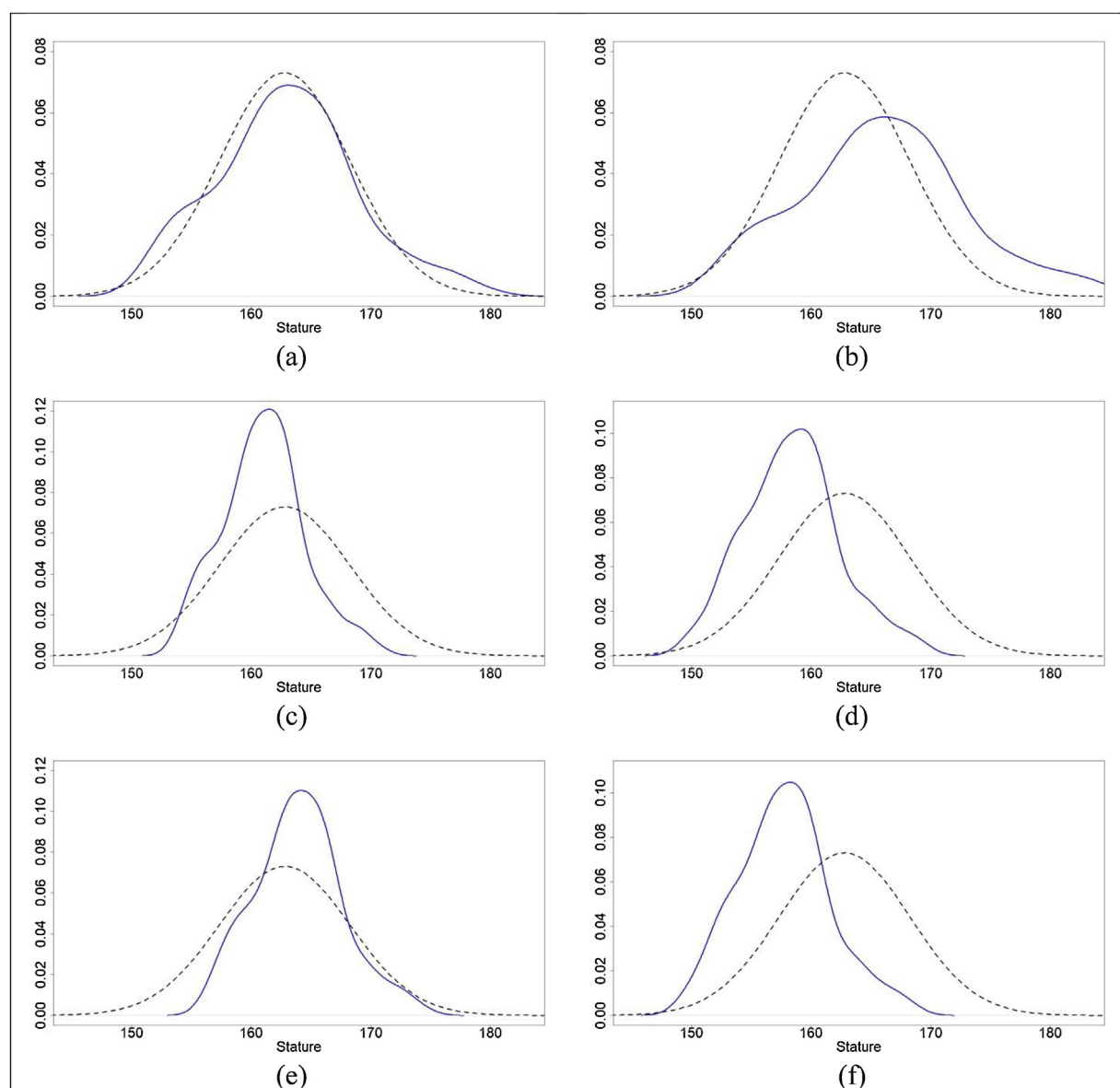[†] Equation of the highest Bayes factor for each type of long bones.

Fig. 2. Graphical comparison of the stature distributions between the Korean War servicemen in Park et al. [36] (dashed line) and stature estimates based on the maximum humeral lengths of the Korean War casualties (solid line). Stature estimates were reconstructed by the equations of (a) Choi et al. [42], (b) Lee et al. [44], (c) Jeong and Jantz [45], (d) Pearson [47], (e) Trotter and Gleser [48], and (f) Fujii [49], respectively. Y-axis indicates density.
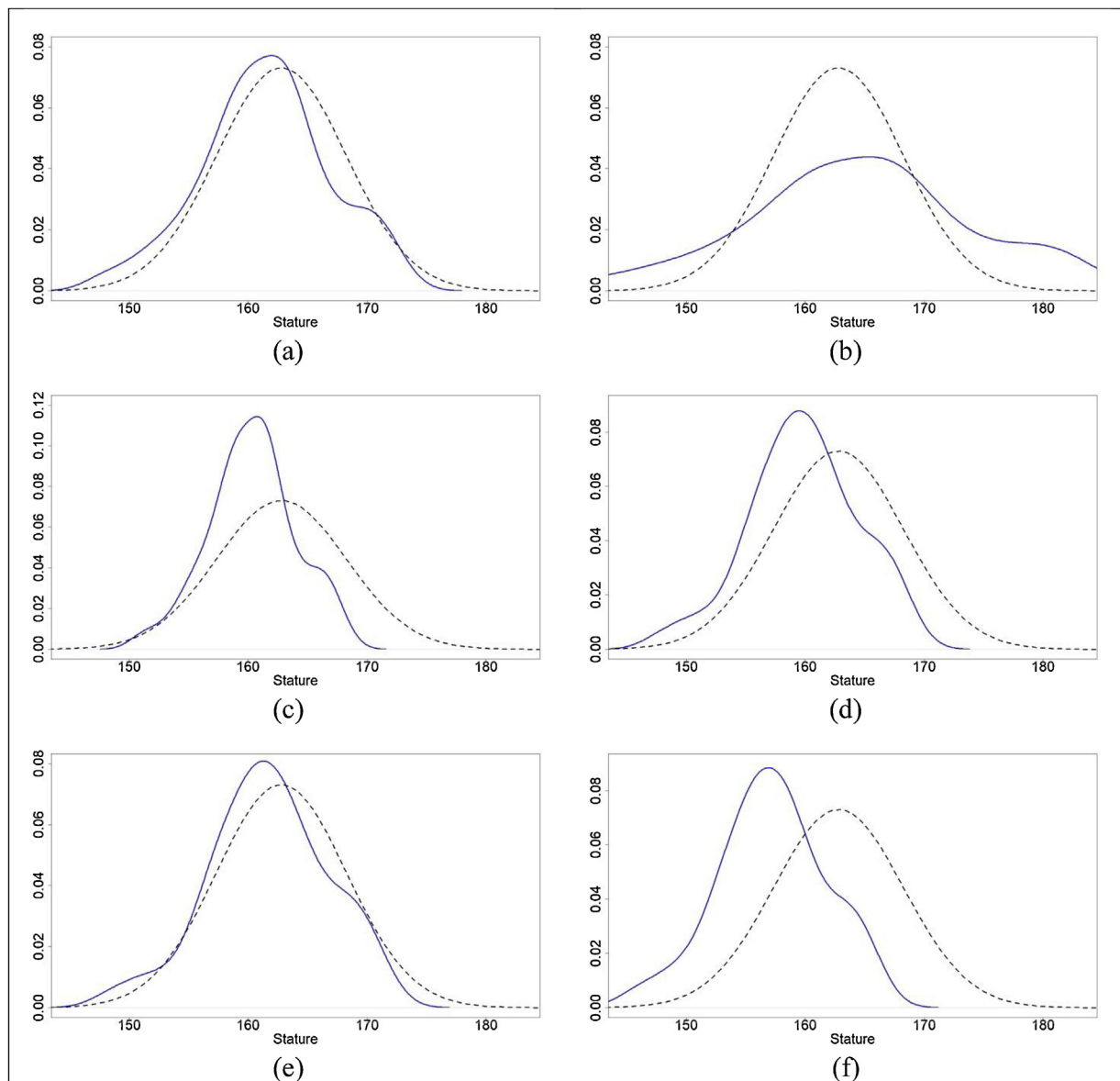
Fig. 3. Graphical comparison of the stature distributions between the Korean War servicemen in Park et al. [36] (dashed line) and stature estimates based on the maximum radial lengths of the Korean War casualties (solid line). Stature estimates were reconstructed by the equations of (a) Choi et al. [42], (b) Lee et al. [44], (c) Jeong and Jantz [45], (d) Pearson [47], (e) Trotter and Gleser [48], and (f) Fujii [49], respectively. Y-axis indicates density.

2.2. Osteometric data of Korean War casualties

Since 2000, MND Agency for KIA Recovery and Identification (MAKRI) of South Korea has performed a nationwide recovery of the Korean War casualty skeletal remains [37]. This study used the osteometric data of the Korean male remains recovered by MAKRI between 2000 and 2006 [38–41]. Of 1400 skeletons recovered during this time period, the maximum length of one or more long bones was reported for 480 individuals (252 individuals with only one bone length and 228 individuals with multiple bone lengths). The maximum femoral length was most frequently reported (n = 397), followed by the tibia (n = 239), humerus (n = 128), radius (n = 49), ulna (n = 34), and fibula (n = 23). The MAKRI reports provide estimated age ranges for each individual whenever available [38–41]. When the medians of the age ranges were regarded as their actual ages, the mean age of the 480 individuals was 22.4 years (SD = 4.5 years).

2.3. Estimation of statures of the Korean War casualties based on their osteometric data

Based on the maximum lengths of the humerus, radius, ulna, femur, tibia, and fibula in the MAKRI reports, the statures of Korean War casualties were reconstructed using different equations which have been generated from Korean samples [42–46] or have been devised from non-Korean samples but often times applied to Korean skeletal remains [47–49]. For the tibia, the spino-malleolar length was regarded as its maximum length following the MAKRI report [38]. Although the tibia lengths in Pearson [47] and Trotter and Gleser [48] did not include the intercondylar eminence, their tibia equations were used without modifications. Since all the six bones were not always included in every study, the number of available equations, and subsequently the number of estimate sets generated by the equations, varied depending on the bones used (Table 1). For example, when the femur was used, seven sets of
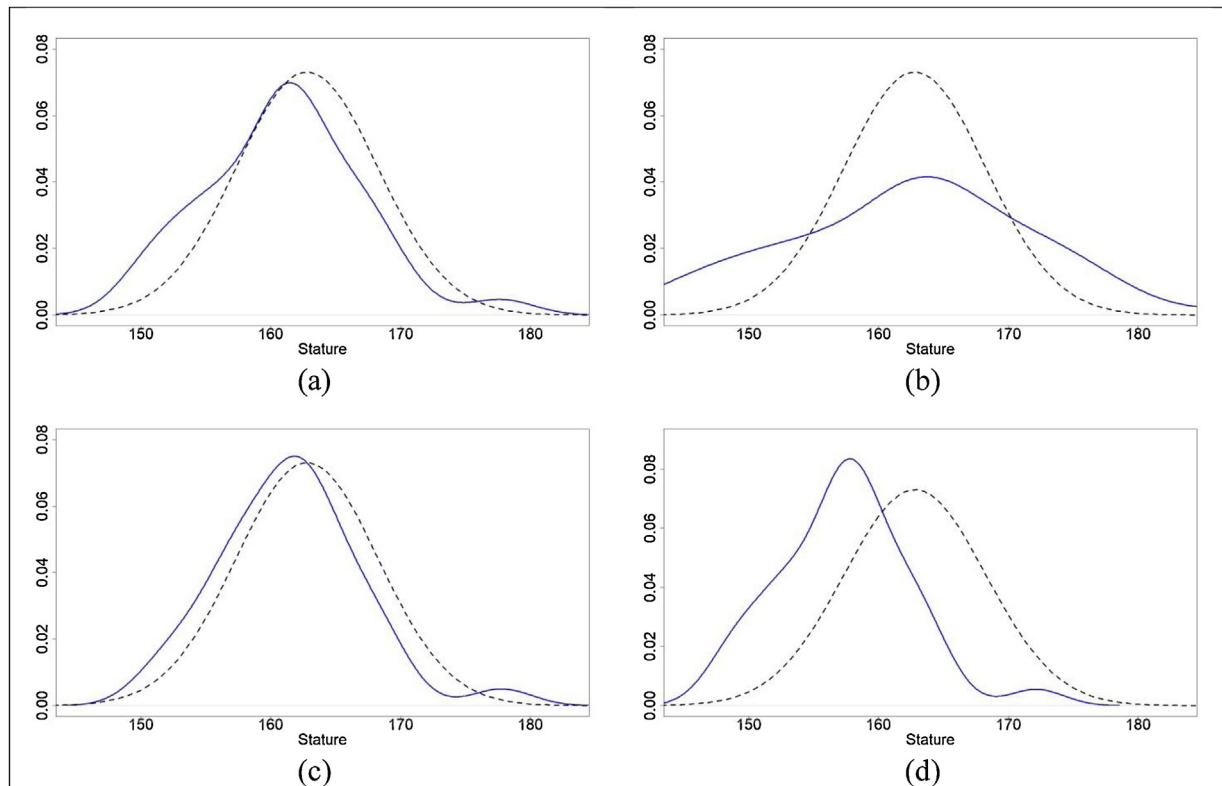
Fig. 4. Graphical comparison of the stature distributions between the Korean War servicemen in Park et al. [36] (dashed line) and stature estimates based on the maximum ulnar lengths of the Korean War casualties (solid line). Stature estimates were reconstructed by the equations of (a) Choi et al. [42], (b) Lee et al. [44], (c) Trotter and Gleser [48], and (d) Fujii [49], respectively. Y-axis indicates density.

stature estimates could be generated by seven different equations, while using the ulna or fibula ended up with only four sets of estimates. When both the left and right sides were present, right bones were preferentially utilized considering that only right bone equations were provided in four out of eight studies [42–44,47]. For the Jeong and Jantz's [45] equations, the mean values of the left and right bones were used following their methodology; however, when only one side of a pair was present, the available side was used. Table 1 shows the equations that were available to each type of long bones.

2.4. Statistical analysis

For each set of stature estimates produced by different equations as well as the stature data provided by Park et al. [36], the normality test was performed using the Kolmogorov-Smirnov test, and also their distributions were visually inspected using Q-Q plots. The normality tests, the paired- and one-sample t-tests as well as calculation of descriptive statistics were carried out using the Statistical Package for the Social Sciences, version 26 (SPSS v.26, IBM Corp., Armonk, NY). All the other analyses were performed in the RStudio version 1.2.5001 for Windows [50]. For a model comparison by the Bayes factors, the LearnBayes package was used, which produced the Bayes factors and posterior probabilities against the null hypothesis [51]. The prior probability was set to 0.5. The R code used in this study is provided in the appendix.

3. Results

Park et al. [36] provided a frequency table of the Korean War servicemen statures between 145 cm and 182 cm with one-centimeter increment. Although decimals were omitted from

their table, a histogram drawn from the table indicates that the stature data does not differ greatly from a normal distribution with skewness and kurtosis of 0.045 and −0.086, respectively (Fig. 1a). The Q-Q plot also shows that most points fall on the straight line with a few extreme ones curved off, which indicates a normal distribution with a slightly heavy tail (Fig. 1b). The Kolmogorov-Smirnov test rejects the null hypothesis that the statures follows a normal distribution, D(12,080) = 4.282, $p < 0.001$; however, this result appears to be simply due to the large sample size which makes kurtosis extremely sensible [52]. Thus, it was assumed that the statures of Korean War servicemen in Park et al. [36] were normally distributed.

Descriptive statistics (mean, standard deviation, min and max values) of the maximum lengths of the left and right long bones in the MAKRI reports are presented in Table 2. The numbers of individuals having both sides with measurable humerus, radius, ulnae, femur, tibia, and fibula were only 21, 6, 1, 85, 45, and 5, respectively. No significant asymmetry was present in the maximum lengths of the left and right humeri, femora, and tibiae (paired t-test; $p = 0.542$, $p = 0.511$, and $p = 0.442$, respectively) (Table 2). The radius, ulna, and fibula were not tested for asymmetry due to a small sample size.

Table 3 shows the descriptive statistics (mean, standard deviation, min and max values) of stature estimates by different equations. The discrepancies between the mean of the Korean War servicemen (162.8 cm) and those of estimates ranged between 0.12 cm (Lee et al.'s [44] ulna equation) and 5.80 cm (Trotter and Gleser's [48] fibula equation). The one sample t-test results indicate that most of the discrepancies are statistically significant at $\alpha = 0.05$ or $\alpha = 0.01$ with exceptions of those by eight equations (Choi et al.'s [42] humerus, radius, ulna, and femur equations; Lee et al.'s [44] radius and ulna equations; and Trotter and Gleser's [48] radius and ulna equations).
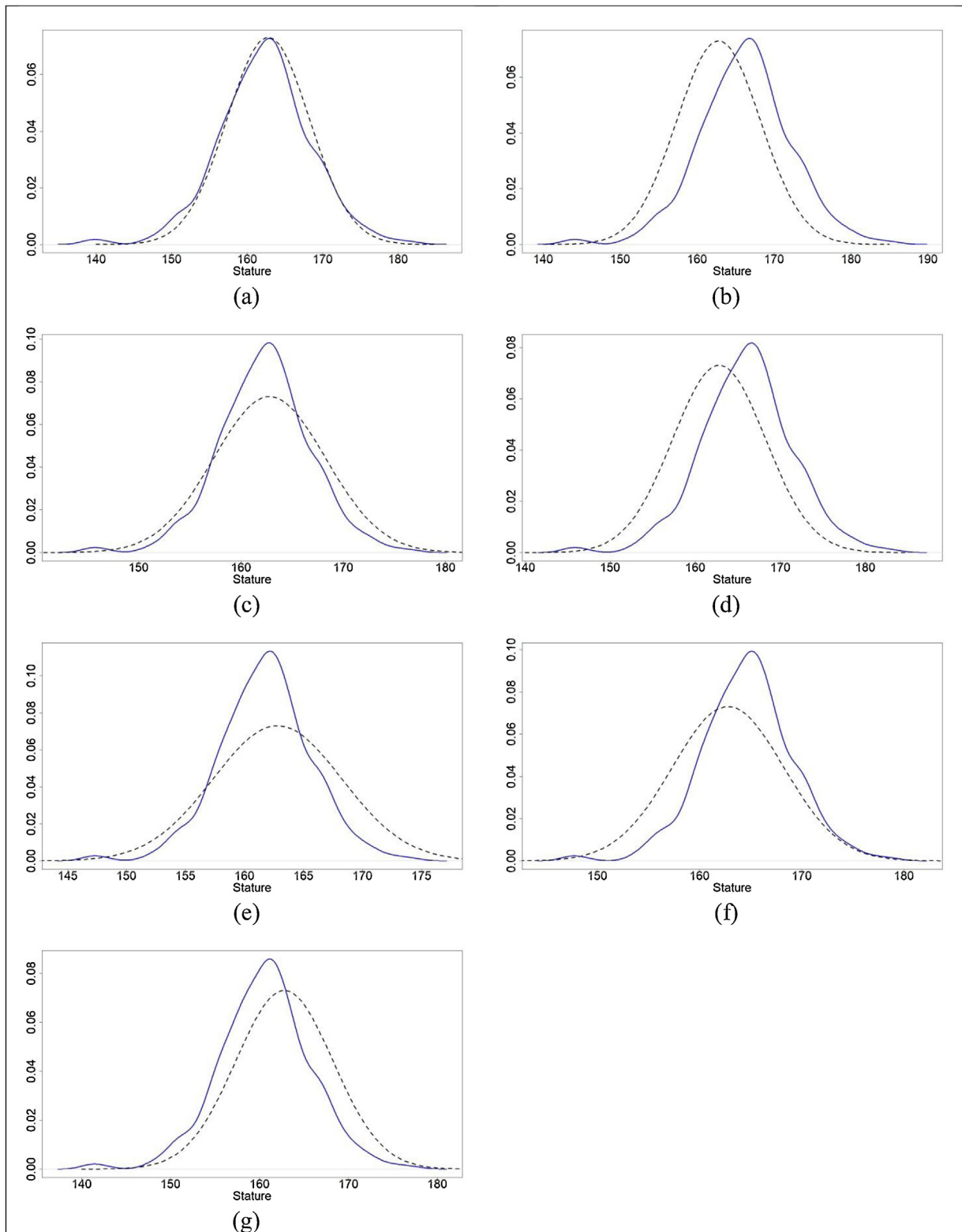
Fig. 5. Graphical comparison of the stature distributions between the Korean War servicemen in Park et al. [36] (dashed line) and stature estimates based on the maximum femoral lengths of the Korean War casualties (solid line). Stature estimates were reconstructed by the equations of (a) Choi et al. [42], (b) Lee et al. [43], (c) Jeong and Jantz [45], (d) Lee et al. [46], (e) Pearson [47], (f) Trotter and Gleser [48], and (g) Fujii [49], respectively. Y-axis indicates density.

The results of Kolmogorov-Smirnov tests indicate that every set of stature estimates are normally distributed at α = 0.05 (Table 4). The normality of the estimates was also confirmed by visual inspections of Q-Q plots (plots not presented here).

Model comparisons between Park et al.'s [36] data and each set of estimates yielded the Bayes factors and posterior probabilities for each equation (Table 5). The highest Bayes factors for the humerus and femur were obtained from Choi et al.'s [42] equations
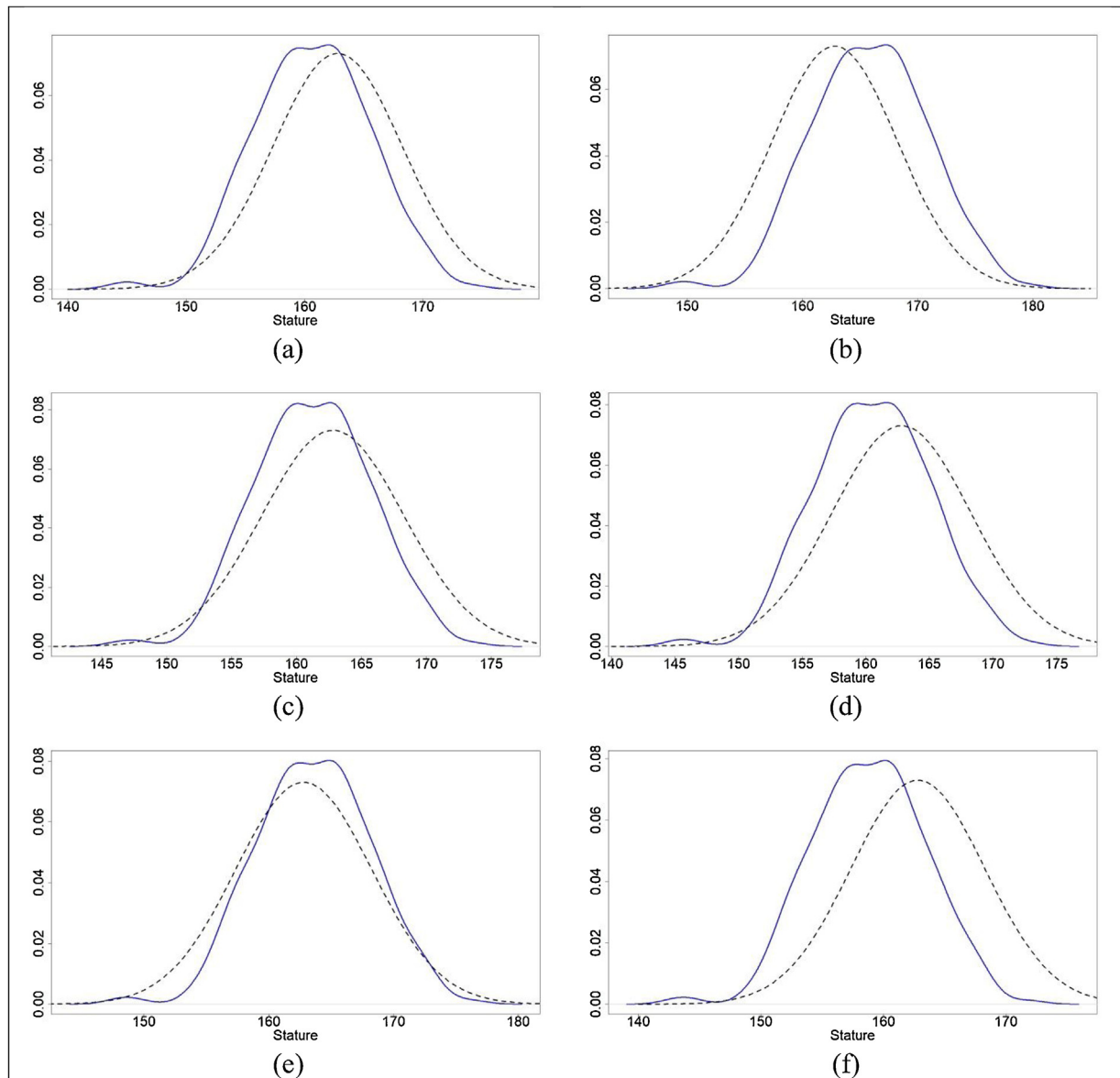
Fig. 6. Graphical comparison of the stature distributions between the Korean War servicemen in Park et al. [36] (dashed line) and stature estimates based on the maximum tibia lengths of the Korean War casualties (solid line). Stature estimates were reconstructed by the equations of (a) Choi et al. [42], (b) Lee et al. [43], (c) Jeong and Jantz [45], (d) Pearson [47], (e) Trotter and Gleser [48], and (f) Fujii [49], respectively. Y-axis indicates density.

(bf = 9.84 and bf = 5.3, respectively); for the radius, ulna, and tibia, from Trotter and Gleser [48] (bf = 2.35, bf = 2.19, and bf = 0.47, respectively); for the fibula, from Lee et al. [43] (bf = 0.53), respectively. It was noted that, for the estimates using the tibia and fibula, even the highest Bayes factors were less than one, and the corresponding posterior probabilities (0.32 and 0.35 for the tibia and fibula, respectively) were less than the prior probability. Graphical comparisons of the distributions are presented in Figs. 2–7.

4. Discussion

Stature of incomplete skeletal remains is reconstructed using the mathematical method which requires a researcher to select a regression equation to apply. General guidelines for a high accuracy of estimates include (i) using an equation generated from a reference sample with the same biological/geographical/temporal backgrounds as a target sample, (ii) using an equation based on a bone dimension (or multiple dimensions) of a highest correlation with stature, and (iii) considering appropriate statistical approaches based on which an equation was originally generated [24,27–30,45,48,53–61]. However, despite these guidelines, subjectivity in the equation selection still remains to some degree, and researchers still look for an objective standard to select the best-performing equation among multiple candidates. Mean difference between population and sample statures has been often used as a quick and simple tool to assess performance of a method (e.g. [42,62]). However, even an identical means between two distributions cannot always be interpreted as an indicator of a good performance of the method. Fig. 8 exemplifies three distributions of the same mean (mean = 7.5) but different shapes: a normal distribution of SD = 2 (blue solid line), a normal distribution of SD = 3 (red dashed line), and bimodal distribution with two peaks at 4.9 and 10.1 (black dotted line). In order to avoid this type of
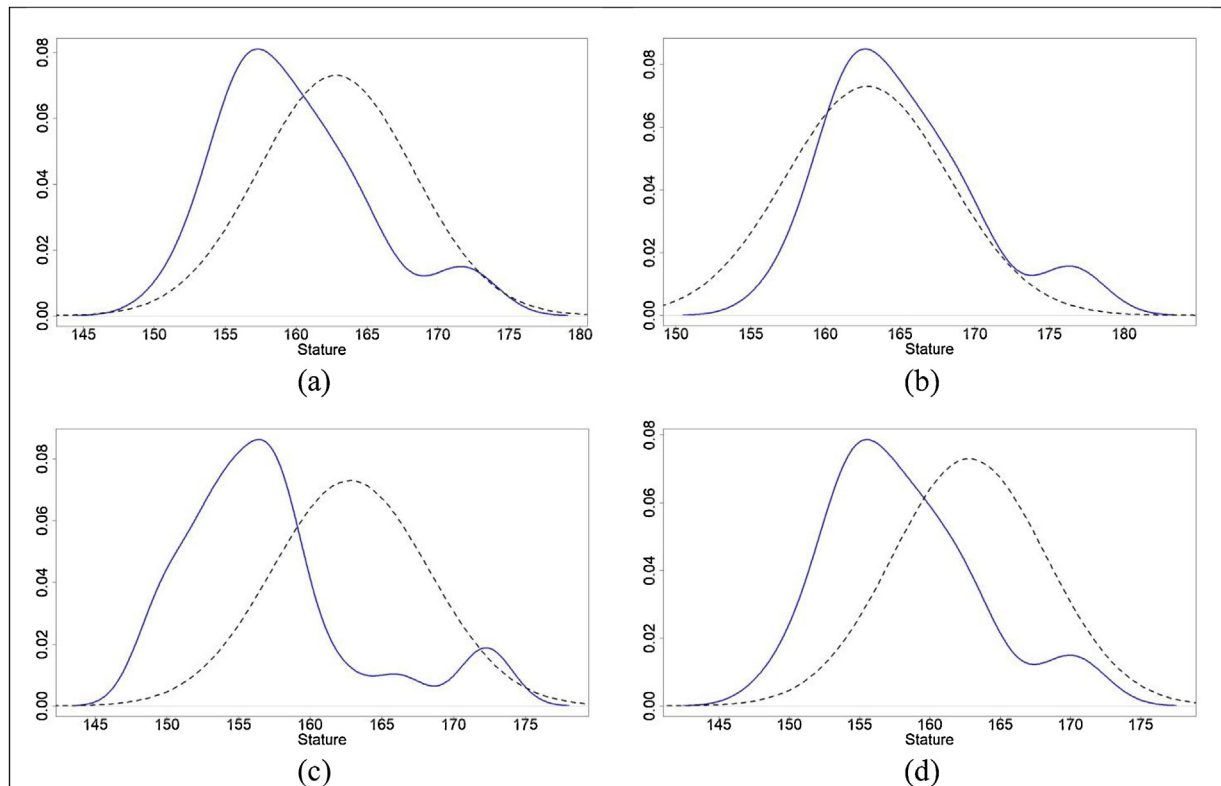
Fig. 7. Graphical comparison of the stature distributions between the Korean War servicemen in Park et al. [36] (dashed line) and stature estimates based on the maximum fibular lengths of the Korean War casualties (solid line). Stature estimates were reconstructed by the equations of (a) Choi et al. [42], (b) Lee et al. [43], (c) Trotter and Gleser [48], and (d) Fujii [49], respectively. Y-axis indicates density.
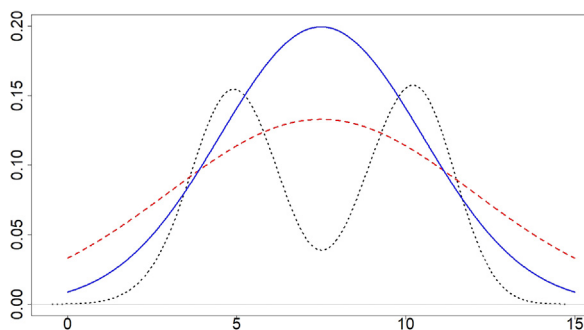

Fig. 8. Exemplary distributions of different shapes but the same mean (mean = 7.5); a normal distribution with SD = 2 (blue solid line), a normal distribution with SD = 3 (red dashed line), and a bimodal distribution with the peaks at 4.9 and 10.1 (black dotted line). Y-axis indicates density (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

misinterpretation and compare performance of different methods in an objective way, a new quantitative standard which can assess the overall shape of distributions is necessary.

Classically, hypothesis testing has typically involved the comparison of two data sets, one of which may be simulated, and a conclusion is reached based on the statistical significance of the relationship of the two sets (e.g., parametric Student's *t*-test and the non-parametric Kolmogorov-Smirnov test). Additionally, multiple models can be tested against a single data set. For statistical model selection, the quality or fitness of various models is quantified as a score that is used to pick the best model. Many criteria for model selection exist, including the Akaike information criterion, the Bayesian information criterion, and the Bayes factor.

In this study, several equations have been proposed that model stature as a function of limited osteometric data. In turn, each equation represents a hypothesis about the underlying distribution, and the appropriateness of each hypothesis can be statistically inferred with model selection. Specifically, thirty-three sets of stature estimates were generated from the osteometric data of the Korean War casualties and the distributions of each set was compared to that of the population (i.e., Korean War servicemen) using the Bayes factors. The Bayesian model comparison allows for quantitative evaluations between two distributions of the same type. Since every set of statue estimates as well as the statures of Korean War servicemen were normally distributed, the goal of the Bayesian model comparison in this study was to compare between two normal distributions with two parameters considered simultaneously (i.e., mean and standard deviation): one from a set of stature estimates and the other from the population.

4.1. Interpretation of Bayes factors

One of the advantages of the Bayes factor over the *p*-value is that it allows for an interpretation about the null model compared to the alternative model. That is, unlike the *p*-value that is simply used to decide whether to 'reject the null or not', the Bayes factor provides an odd to indicating how much better the null (or alternative) model is than the alternative (or null) model [63]. For example, if the Bayes factor of the null model (M_0) over the alternative (M_1) is 0.5 (i.e., $BF_{01} = 0.5$), it suggests that the given data are 0.5 times more likely to occur under the null model than the alternative model. At the same time, its inverse value (i.e., $1/BF_{01} = 2$) allows for an interpretation in terms of the alternative model, which is that the data are twice more likely to occur under the null model than the null model.

Table 6
Interpretation of Bayes factors (BF$_{01}$) with evidence categories (adapted from Raftery [64] and Jeffreys [65] with modifications).

Bayes factor	Interpretation	
	Raftery [64]	Jeffreys [65]
>150	Decisive evidence for M$_0$	Decisive evidence for M$_0$
100–150	Strong evidence for M$_0$	Decisive evidence for M$_0$
30–100	Strong evidence for M$_0$	Very strong evidence for M$_0$
20–30	Strong evidence for M$_0$	Strong evidence for M$_0$
10–20	Positive evidence for M$_0$	Strong evidence for M$_0$
3–10	Positive evidence for M$_0$	Substantial evidence for M$_0$
1–3	Weak evidence for M$_0$	Anecdotal* evidence for M$_0$
1	No evidence	No evidence
1/3–1	Weak evidence for M$_1$	Anecdotal* evidence for M$_1$
1/10–1/3	Positive evidence for M$_1$	Substantial evidence for M$_1$
1/20–1/10	Positive evidence for M$_1$	Strong evidence for M$_1$
1/30–1/20	Strong evidence for M$_1$	Strong evidence for M$_1$
1/100–1/30	Strong evidence for M$_1$	Very strong evidence for M$_1$
1/150–1/100	Strong evidence for M$_1$	Decisive evidence for M$_1$
<1/150	Decisive evidence for M$_1$	Decisive evidence for M$_1$

* 'Anecdotal' is a replacement of 'worth no more than a bare mention' following Wetzels et al. [66].

In this study, the parameters (mean and standard deviation) of the stature distributions from the population (i.e., Korean War servicemen) and estimated statures were set as the null (M$_0$) and alternative models (M$_1$), respectively. Thus, a high Bayes factor (BF$_{01}$) indicates a high likelihood of the scenario that the estimated statures occur from the distribution of the population or a small discrepancy of the two distributions. Out of 33 sets of stature estimates, the highest Bayes factor was obtained from those produced by Choi et al.'s [42] humerus equation (BF = 9.84) (Table 5), which indicates that this equation performs best among all the equations under comparison in this study.

Some guidelines have been suggested for interpretation of the Bayes factor by previous researchers. Following the guidelines in Table 6, the Bayes factors from Choi et al.'s [42] humerus and femur equations (BF = 9.84 and 5.3, respectively) can be a positive or substantial evidence that the estimated statures by those equations follow the distribution of the population statures. All the other equations ended up with the Bayes factors less than three, which implies that the null model can only be supported to a weak/anecdotal extent or cannot be supported at all.

For eight out of 33 equations under comparison in this study (Choi et al.'s [42] humerus, radius, ulna, and femur equations; Lee et al.'s [44] radius and ulna equations; and Trotter and Gleser's [48] radius and ulna equations), the results of the one sample *t*-tests could not reject the null hypothesis that the difference between the true mean of the stature estimates and population mean is zero (Table 3). However, the Bayes factors from six of them (radius and ulna equations of Choi et al. [42], Lee et al. [44], and Trotter and Gleser [48]) do not provide a positive or substantial evidence supporting that their estimates re-produce the distribution of the population. Thus, given all other conditions being equal, these six equations should be only used with caution or should not be prioritized than the other two equations of the Bayes factors greater than three. It was also noted that, for the stature estimates produced by Jeong and Jantz's [45] femur equation, the null hypothesis of the one sample *t*-test was rejected, while the corresponding Bayes factor was relatively high (bf = 2.7; posterior probability = 0.73). In fact, the Bayes factor of 2.7 is the third greatest value of the 33 values generated in this study. It is often pointed out that, given two studies of different sample sizes, the same *p*-values do not represent the same extent of statistical weight of evidence since even a small difference tends to result in a

'significant difference between groups' in a large sample study [53,67]. In this study, the reason why the one sample *t*-test rejected the null hypothesis regarding the Jeong and Jantz' [45] femur equation is also more likely due to a large sample size (n = 397) rather than its poor performance. On the other hand, the Bayes factors are comparable to each other regardless of samples size since it just provides the odds ratio between two probabilities. These results clearly raise not only a possible demerit when the *p*-value is utilized for a comparison purpose but also a usefulness of the Bayes factors for evaluating and prioritizing multiple equations of interest.

It is beyond the scope of this study to investigate specific reasons for the high or low Bayes factors associated with each equation under comparison. Hence, additional research on the factors that likely affected performance of each equation (e.g., discrepancy in the geographic, temporal, or biological backgrounds between the reference and target sample) will be necessary in the future. However, the results of this study clearly demonstrated that higher Bayes factors could be obtained when two stature distributions from the population and sample were largely similar to each other (Table 5, Figs. 2–7). For example, the distributions of the stature estimates produced by Choi et al.'s [42] humerus and femur equations (bf = 9.84 and bf = 5.3, respectively) almost overlap with that of the population (Figs. 2a and 5a). Thus, in cases where multiple equations are available, researchers need to compare the Bayes factors from each equation and select the one with the highest Bayes factor. Also, even when there is only one equation available, researchers need to calculate the Bayes factor from the equation and assess its performance based on the guidelines presented in Table 6.

The Bayesian model comparison requires known prior information which will not always be available, particularly in the skeletal assemblage of the past. In this situation, if there are a number of complete skeletons in the assemblage, researchers can consider reconstructing statures of the complete skeletons using the anatomical method to build prior information. In fact, stature estimates based on the anatomical method, particularly the revised Fully technique [23], are regarded highly accurate and have often served as 'true statures' in the hybrid method [e.g., 23,45,68]. Thus, it appears also reasonable to use the anatomically reconstructed statures for prior information in the model selection procedure using the Bayes factor, which will subsequently extend the utility of the Bayes factors to target samples with no known information.

CRediT authorship contribution statement

Yangseung Jeong: Project administration, Conceptualization, Data curation, Methodology, Writing - original draft. **Ashlin P. Harris:** Methodology, Software, Writing - original draft, Writing - review & editing. **Omar Ali:** Validation, Writing - review & editing. **Yochun Jung:** Conceptualization, Supervision, Methodology, Validation, Investigation, Resources, Writing - review & editing.

Acknowledgements

Appendix A

R codes for comparing methods using the Bayes factor.

```
# An example input data saved as a csv file:
"ID","Method_A","Method_B","Method_C"
"01",  146.23,   150.60,   147.69
"02",  167.30,   172.22,   172.11
"03",  165.50,   160.35,   164.74
"04",  169.25,   182.62,   180.13


# Read in stature estimates
stature<-read.table("example.csv",header=TRUE,sep=",",row.names=1)


# Calculate means, counts, and standard deviations
data<-rbind(
   colMeans(stature,na.rm=TRUE),
   apply(stature,2,function(x) length(x[!is.na(x)])),
   apply(stature,2,sd,na.rm=TRUE)
                )
row.names(data)<-c('Mean','Count','Std.Dev')


# Calculate Bayes factors and posterior probabilities
mu<-162.11  # Population mean
sigma<-4.33 # Standard deviation
library(LearnBayes)
Bayesian<-apply(data,2,function(x) mnormt.twosided(mu,prob<-0.5,sigma,x))


# Print results
results<-matrix(unlist(Bayesian),nrow=2)
colnames(results)<-names(stature)
rownames(results)<-c('bf','post')
print(t(results))
```

References

[1] S. Ulijaszek, J. Komlos, From a history of anthropometry to anthrometric history, in: C.G.N. Mascie-Taylor, A. Yasukouchi, S. Ulijaszek (Eds.), Human Variation: From the Laboratory to the Field, CRC Press, 2010, pp. 183–197.

[2] R.M. Malina, M.E. Reyes, B.B. Little, Secular change in heights of indigenous adults from a Zapotec-speaking community in Oaxaca, southern Mexico, Am. J. Phys. Anthropol. 141 (3) (2010) 463–475.

[3] M.N. Cohen, G.M.M. Crane-Kramer, Ancient Health: Skeletal Indicators of Agricultural and Economic Intensification, University Press of Florida, Gainesville, 2007.

[4] K.R. Rosenberg, L. Zune, C.B. Ruff, Body size, body proportions, and encephalization in a Middle Pleistocene archaic human from northern China, Proc. Natl. Acad. Sci. U. S. A. 103 (10) (2006) 3552–3556.

[5] K.J. Weinstein, Body proportions in ancient Andeans from high and low altitudes, Am. J. Phys. Anthropol. 128 (3) (2005) 569–585.

[6] R.H. Steckel, J.C. Rose, The Backbone of History: Health and Nutrition in the Western Hemisphere, Cambridge University Press, Cambridge, 2002.

[7] C.B. Ruff, Variation in human body size and shape, Annu. Rev. Anthropol. 31 (2002) 211–232.

[8] B. Bogin, R. Keep, Eight thousand years of economic and political history in Latin America revealed by anthropometry, Ann. Hum. Biol. 26 (4) (1999) 333–351.

[9] C.B. Ruff, Morphological adaptation to climate in modern and fossil hominids, Yearb. Phys. Anthropol. 37 (1994) 65–107.

[10] W.L. Jungers, Size and Scaling in Primate Biology, Plenum Press, New York, 1985.

[11] P. Smith, L.K. Horowitz, Archaeological and skeletal evidence for dietary change during the late Pleistocene/Early holocene in the levant, in: M.N. Cohen, G.J. Armelagos (Eds.), Paleopathology at the Origin of Agriculture, Academic Press, Orlando, 1984, pp. 101–136.

[12] K. Schmidt-Nielson, Scaling: Why is Animal Size so Important? Cambridge University Press, Cambridge, 1984.

[13] D.W. Frayer, Biological and cultural change in the European late pleistocene and early holocene, in: F.H. Smith, F. Spencer (Eds.), The Origins of Modern Humans: A World Survey of the Fossil Evidence, Wiley-Liss, New York, 1984, pp. 1–50.

[14] W.A. Calder, Size, Function, and Life History, Harvard University Press, Cambridge, Mass, 1984.

[15] R.H. Peters, The Ecological Implications of Body Size., Cambridge Cambridge-shire, Cambridge University Press, New York, 1983.

[16] T.A. McMahon, J.T. Bonner, On Size and Life, Scientific American Library: Distributed by W.H. Freeman, New York, (1983) .

[17] P.B. Eveleth, J.M. Tanner, Worldwide Variation in Human Growth, Cambridge University Press, Cambridge, UK, 1976.

[18] Y. Jeong, G. Jeong, S. Pergande, K. Lee, Generating a vertebrae-based method to discriminate between Koreans and U.S. white male casualties from the Korean war, J. Forensic Sci. 64 (6) (2019) 1776–1781.

[19] I.C. Hwang, K.K. Kim, H.C. Kang, D.R. Kang, Validity of stature-predicted equations using knee height for elderly and mobility impaired persons in Koreans, Epidemiol. Health 31 (2009) e2009004.

[20] A. Özaslan, M.Y. İşcan, İ. Özaslan, H. Tuğcu, S. Koç, Estimation of stature from body parts, Forensic Sci. Int. 132 (1) (2003) 40–45.

[21] P. Uhrová, R. Beňuš, S. Masnicová, Stature estimation from various foot dimensions among Slovak population, J. Forensic Sci. 58 (2) (2013) 448–451.

[22] P. Uhrová, R. Beňuš, S. Masnicová, Z. Obertová, D. Kramárová, K. Kyselicová, M. Dörnhöferová, S. Bodoriková, E. Neščáková, Estimation of stature using hand and foot dimensions in Slovak adults, Leg. Med. 17 (2) (2015) 92–977.

[23] M.H. Raxter, B.M. Auerbach, C.B. Ruff, Revision of the fully technique for estimating statures, Am. J. Phys. Anthropol. 130 (3) (2006) 374–384.

[24] C.B. Ruff, B.M. Holt, M. Niskanen, V. Sladek, M. Berner, E. Garofalo, H.M. Garvin, M. Hora, H. Maijanen, S. Niinimaki, K. Salo, D. Schuplerová, Tompkins, stature and body mass estimation from skeletal remains in the European Holocene, Am. J. Phys. Anthropol. 148 (4) (2012) 601–617.

[25] H. Maijanen, Testing anatomical methods for stature estimation on individuals from the W.M. Bass Donated Skeletal Collection, J. Forensic Sci. 54 (4) (2009) 746–752.

[26] M.H. Raxter, C.B. Ruff, B.M. Auerbach, Technical note: revised fully stature estimation technique, Am. J. Phys. Anthropol. 133 (2) (2007) 817–818.

[27] Y. Jeong, E.J. Woo, Analytical review of the forensic anthropological techniques for stature estimation in Korea, korean, J. Phys. Anthropol. 31 (4) (2018) 121–131.

[28] Y. Jeong, E.J. Woo, Estimation of stature in archaeological human skeletal remains, Field Archaeol. 35 (2019) 129–154.

[29] M. Trotter, G.C. Gleser, Trends in stature of american whites and negroes born between 1840 and 1924, Am. J. Phys. Anthropol. 9 (4) (1951) 427–440.

[30] M. Trotter, G.C. Gleser, Estimation of stature from long bones of American Whites and Negroes, Am. J. Phys. Anthropol. 10 (4) (1952) 463–514.

[31] V. Formicola, More is not always better: trotter and Gleser's equations and stature estimates of upper paleolithic european samples, J. Hum. Evol. 45 (3) (2003) 239–244.

[32] T.S. Constandse-Westermann, M.L. Blok, R.R. Newell, Long bone length and stature in the western european mesolithic. I. Methodological problems and solutions, J. Hum. Evol. 14 (4) (1985) 399–410.

[33] J. Albert, Bayesian Computation with R, 2nd ed., Springer Science & Business Media, New York, 2009.

[34] Z. Dienes, Using Bayes to get the most out of nonsignificant results, Front Psychol. 5 (2014) 781, doi:http://dx.doi.org/10.3389/fpsyg.2014.00781.

[35] A.E. Raftery, Hypothesis testing and model selection, in: W.R. Gilks, D.J. Spiegelhalter, S. Richardson (Eds.), Markov Chain Monte Carlo in Practice, Chapman and Hall, London, 1996, pp. 163–188.

[36] T. Park, H. Choung, M. Lee, S. Chang, Anthropological studies on the Korean: I. Pro-standard of the length, weight and girth of the chest of recruit, Medicine 1 (1953) 107–112.

[37] MND Agency for KIA Recovery and Identification, (2019) http://www.withcountry.mil.kr/cop/withcountry/withExcavateView.do?id=withcountry_020201000000. (Accessed 18 November 2019).

[38] ROK Army Headquarters, Institute of Chungwon Culture, In Commemorating the 50th Anniversary of the Korean War: Excavation Report of the Casualties from the Korean War, Gippumsa Press, Seoul, 2003.

[39] ROK Army Headquarters, Institute of Chungwon Culture, 2004 Excavation Report of the Casualties from the Korean War, Hakyeon Press, Chungbuk, 2004.

[40] ROK Army Headquarters, Institute of Chungwon Culture, The 2005 Excavation Report of the KIA (Killed in Action) Remains from the Korean War, Ilkwang Press, Chungbuk, 2005.

[41] ROK Army Headquarters, Institute of Chungwon Culture, The 2006 Excavation Report of the KIA (Killed in Action) Remains from the Korean War, Ilkwang Presss, Chungbuk, 2006.

[42] B. Choi, Y. Chae, I. Chung, H. Kang, Correlation between the postmortem stature and the dried limb-bone lengths of Korean adult males, Yonsei Med. J. 38 (2) (1997) 79–85.

[43] J.H. Lee, Y.S. Kim, U.Y. Lee, D.K. Park, Y.K. Jeong, N.S. Lee, S.Y. Han, S.H. Han, Stature estimation from partial measurements and maximum length of lower limb bones in Koreans, Aust. J. Forensic Sci. 46 (3) (2014) 330–338.

[44] J.H. Lee, Y.G. Jeong, N.S. Lee, S.Y. Han, S.H. Han, Stature estimation from partial measurements and maximum length of upper limb bones in Koreans, Int. J. Res.-GRANTHAALAYAH 3 (10) (2015) 100–107.

[45] Y. Jeong, L.M. Jantz, Developing Korean-specific equations of stature estimation, Forensic Sci. Int. 260 (2016) 105.e1–105.e11.

[46] S. Lee, H.H. Gong, J.Y. Hyun, H.N. Koo, H.Y. Lee, N.E. Chung, Y.S. Choi, K.M.Y, B.H. Choi, Estimation of stature from femur length measured using computed tomography after the analysis of three-dimensional characteristics of femur bone in Korean cadavers, Int. J. Legal Med. 131 (2017) 1355–1362.

[47] K. Pearson, Mathematical contributions to the theory of evolution. V. On the reconstruction of the stature of prehistoric races, Philos. T. R. Soc. S-A 192 (1899) 169–244.

[48] M. Trotter, G.C. Gleser, A re-evaluation of estimation of stature based on measurements of stature taken during life and of long bones after death, Am. J. Phys. Anthropol. 16 (1) (1958) 79–123.

[49] A. Fujii, On the relation of long bone lengths of limbs to stature, Juntendodaigaku Taiikugakubu Kiyo 3 (1960) 49–61.

[50] R Core Team, R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2017. https://www.R-project.org/.

[51] J. Albert, LearnBayes: Functions for Learning Bayesian Inference, R Package Version 2.15.1, (2018) . https://CRAN.R-project.org/package=LearnBayes/.

[52] A. Field, Discovering Statistics Using SPSS, 3rd ed., SAGE Publication Ltd., London, 2009, pp. 131–165.

[53] Y. Jeong, L.M. Jantz, Caveats in using Trotter and Gleser's (1958) Asian equations for stature estimation, Korean J. Phys. Anthropol. 29 (3) (2016) 81–91.

[54] W.M. Krogman, The Human Skeleton in Forensic Medicine, Springfield, Illinois, 1962.

[55] V. Formicola, M. Franceschi, Regression equations for estimating stature from long bones of early holocene european samples, Am. J. Phys. Anthropol. 100 (1) (1996) 83–88.

[56] T. Sjøvold, Estimation of stature from long bones utilizing the line of organic correlation, Hum. Evol. 5 (5) (1990) 431–447.

[57] J.K. Lundy, The mathematical versus anatomical methods of stature estimates from long bones, Am. J. Forensic Med. Pathol. 6 (1) (1985) 73–76.

[58] P.H. Stevenson, On racial differences in stature long bone regression formulae with special reference to stature reconstruction formulae for the Chinese, Biometrika 21 (1) (1929) 303–321.

[59] I. Duyar, C. Pelin, R. Zagyapan, A new method of stature estimation for forensic anthropological application, Anthropol. Sci. 114 (1) (2006) 23–27.

[60] AAFS Standard Board, Standard for Stature Estimation in Forensic Anthropology, (2019) www.asbstandardsboard.org. (Accessed 19 November 2019).

[61] L.W. Konigsberg, S.M. Hens, L.M. Jantz, W.L. Jungers, Stature estimation and calibration: Bayesian and maximum likelihood perspectives in physical anthropology, Am. J. Phys. Anthropol. Suppl. 27 (1998) 65–92.

[62] B.M. Auerbach, Methods for estimating missing human skeletal element osteometric dimensions employed in the revised fully technique for estimating stature, Am. J. Phys. Anthropol. 145 (1) (2011) 67–80.

[63] A.F. Jarosz, J. Wiley, What are the odds? A practical guide to computing and reporting Bayes factors, J. Probl. Solving 7 (1) (2014) 1–9, doi:http://dx.doi.org/10.7771/1932-6246.1167.

[64] A.E. Raftery, Bayesian model selection in social research, in: P.V. Marsden (Ed.), Sociological Methodology 1995, Blackwell, Cambridge, MA, 1995, pp. 111–196.

[65] H. Jeffreys, Theory of Probability, 3rd ed., Oxford University Press, Oxford, UK, 1961.

[66] R. Wetzels, D. Matzke, M.D. Lee, J.N. Rouder, G.J. Iverson, E.J. Wagenmakers, Statistical evidence in experimental psychology: an empirical comparison using 855 t tests, Perspect. Psychol. Sci. 6 (3) (2011) 291–298, doi:http://dx.doi.org/10.1177/1745691611406923.

[67] E.J. Wagenmakers, A practical solution to the pervasive problems of p values, Psychon. B Rev. 14 (5) (2007) 779–804.

[68] B.M. Auerbach, C.B. Ruff, Stature estimation formulae for indigenous populations from North America, Am. J. Phys. Anthropol. 141 (2) (2010) 190–207.