10-2022

Exact Statistical Distribution and Correlation of Human Height and Weight: Analysis and Experimental Confirmation

Mark P. Silverman
Trinity College, mark.silverman@trincoll.edu

Scientific Research Publishing

Open Journal of Statistics, 2022, 12, 743-787
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Exact Statistical Distribution and Correlation of Human Height and Weight: Analysis and Experimental Confirmation

Mark P. Silverman

Department of Physics, Trinity College, Hartford, USA
Email: mark.silverman@trincoll.edu

How to cite this paper: Silverman, M.P. (2022) Exact Statistical Distribution and Correlation of Human Height and Weight: Analysis and Experimental Confirmation. *Open Journal of Statistics*, **12**, 743-787.
https://doi.org/10.4236/ojs.2022.125044

Received: August 28, 2022
Accepted: October 28, 2022
Published: October 31, 2022

Abstract

The statistical relationship between human height and weight is of especial importance to clinical medicine, epidemiology, and the biology of human development. Yet, after more than a century of anthropometric measurements and analyses, there has been no consensus on this relationship. The purpose of this article is to provide a definitive statistical distribution function from which all desired statistics (probabilities, moments, and correlation functions) can be determined. The statistical analysis reported in this article provides strong evidence that height and weight in a diverse population of healthy adults constitute correlated bivariate lognormal random variables. This conclusion is supported by a battery of independent tests comparing empirical values of 1) probability density patterns, 2) linear and higher order correlation coefficients, 3) statistical and hyperstatistics moments up to 6th order, and 4) distance correlation (dCor) values to corresponding theoretical quantities: 1) predicted by the lognormal distribution and 2) simulated by use of appropriate random number generators. Furthermore, calculation of the conditional expectation of weight, given height, yields a theoretical power law that specifies conditions under which body mass index (BMI) can be a valid proxy of obesity. The consistency of the empirical data from a large, diverse anthropometric survey partitioned by gender with the predictions of a correlated bivariate lognormal distribution was found to be so extensive and close as to suggest that this outcome is not coincidental or approximate, but may be a consequence of some underlying biophysical mechanism.

Keywords

Correlation of Height and Weight, Distribution of Height and Weight, Body Mass Index, Lognormal Distribution, Distance Correlation (dCor), Hyperstatistics

1. Introduction

Scientific interest in the values and correlations of anthropometric data trace back to the beginnings of modern statistics in the late 19th and early 20th Centuries with the researches of Quetelet, Galton, Pearson, and others [1] [2] [3]. These many studies established the Gaussian function as the mathematical expression best approximating the distribution of such human features as height, weight, and other biometric attributes. So pervasive has been the Gaussian distribution that it is ubiquitously referred to as the "normal distribution", a reference probably dating back to Quetelet's influential study of "the average man" (L'homme Moyen) in 1835 [4].

Although more refined studies have revealed that anthropometric data can show deviations from normality, attempts to find relationships between human height and weight remained uncertain, controversial and based on approximate or indirect methods such as data fitting [5], mechanical modeling [6], and gene identification [7]. Height and weight are of particular importance since they directly relate to the body mass index (BMI) [8], which is a measure of obesity and a risk factor for metabolic disease [9] and Alzheimer's Disease [10]. A previous paper by Silverman and Lipscombe [11], to be referred to as Part I, determined the mathematically exact statistical distribution of BMI.

The present paper, to be regarded as Part II, provides evidence for the proposition that, in a healthy adult human population with access to adequate nutrition, height and weight are distributed as correlated bivariate lognormal random variables. This conclusion is supported by a comprehensive investigation comprising four independent components:

1) Tests of the correlation functions of the height and weight of a large anthropometric data set of individuals, partitioned by gender, against predictions of the bivariate lognormal distribution;

2) Search for nonlinear correlations, *not* attributable to the bivariate lognormal distribution, by means of a sensitive nonparametric algorithm known as distance correlation [12] [13];

3) Comparison of statistical tests of the empirical anthropometric data set with identical tests performed on comparably sized populations artificially created with correlated lognormal random number generators (RNGs);

4) Tests of the marginal distributions of height and weight against predictions of associated univariate lognormal distributions, and of the natural logarithms of height and weight against predictions of associated univariate normal distributions.

The outcome of this four-part analysis shows that linear and higher-order correlations of human height and weight are *predictable* in terms of the *single* Pearson correlation coefficient for height and weight employed in the lognormal probability density function (PDF). Moreover, agreement between the empirical data and the predictions of lognormal theory is so extensive as to suggest that the lognormal distribution of adult human height and weight is not approx-

imate, but an exact distribution possibly characteristic of a more fundamental underlying biophysical mechanism.

1.1. Marginal Distributions of Height and Weight

Part I [11] reported the exact probability density function of BMI that follows mathematically from the defining relation

$$B = W/H^2 \tag{1}$$

in which height H is expressed in meters (m), the corresponding weight W is expressed in the mass unit kilograms (kg), and B is the BMI expressed in kg/m^2. It is to be stressed that H, W, and therefore B, are random variables, which means that information and interpretations extrapolated from the BMI PDF refer to *populations*, and not to individuals, an essential point not always understood by the lay news media [11] [14].

The specific form that the general BMI density function takes depends on the statistical distributions of H and W. Such empirical distributions are often represented visually as histograms. However, if two random variables are not independent, then the histogram of each is a graphical representation of the *marginal* distribution of that variable, and provides no information regarding the correlation of the two variables. In Part I evidence was provided to show that height and weight of individuals measured in the Anthropometric Survey of U.S. Army Personnel (ANSUR)—a large data base comprising 4082 males and 1986 females [15]—were highly correlated. **Figure 1** shows scatter plots of W against H for the separate male and female cohorts. The two patterns suggest a significant linear correlation. The superposed curves, to be discussed later, are the lines of regression (dashed red) and the conditional expectation functions (solid blue). Descriptive statistics are given in **Table 1** for the two cohorts, together with theoretically predicted values, where appropriate. Details of **Table 1** will be discussed at relevant points throughout the paper.

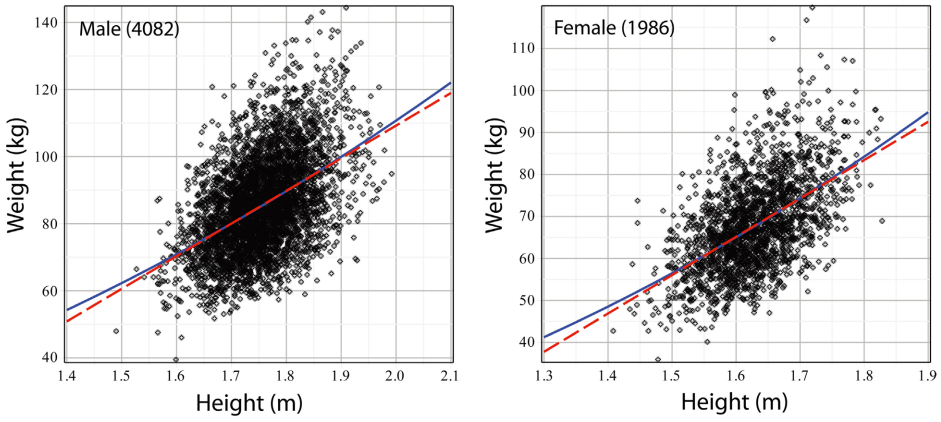


Figure 1. Correlation of weight and height for males (left) and females (right) of the ANSUR population. Lines of regression (dashed red) are obtained by the method of least squares. The conditional expectation functions of weight given height (solid blue) are calculated from the lognormal PDF (17), using ANSUR parameters in **Table 1**.

Table 1. Descriptive statistics of height and weight of ANSUR population.

Marginal Statistic	**Cohort M: 4082 F: 1986**	**Height (m) Empirical**	**Height (m) Theory**	**Weight (kg) Empirical**	**Weight (kg) Theory**
Mean	Male	1.7562	1.7562	85.5240	85.5224
	Female	1.6285	1.6285	67.7582	67.7527
Standard Deviation	Male	0.0685	0.0685	14.2190	14.2427
	Female	0.0642	0.0642	10.9819	10.9378
Skewness Sk	Male	0.1113	0.1171	0.4817	0.5042
Standard Error SE_{Sk}	Male	-	0.0383	-	0.0383
Skewness Sk	Female	0.0876	0.1183	0.5545	0.4885
Standard Error SE_{Sk}	Female	-	0.0549	-	0.0549
Kurtosis K	Male	3.0680	3.0244	3.3583	3.4554
Standard Error SE_K	Male	-	0.0766	-	0.0766
Kurtosis K	Female	3.0040	3.0249	3.6599	3.4273
Standard Error SE_K	Female	-	0.1098	-	0.1098
Marginal Statistic	**Cohort M: 4082 F: 1986**	**LnHeight Empirical**	**LnHeight Theory**	**LnWeight Empirical**	**LnWeight Theory**
Mean	Male	m_H 0.5624	-	m_W 4.4351	-
	Female	m_H 0.4869	-	m_W 4.2030	-
Standard Deviation	Male	s_H 0.0390	-	s_W 0.1654	-
	Female	s_H 0.0394	-	s_W 0.1604	-
Skewness Sk	Male	−0.0090	0	-0.0193	0
Standard Error SE_{Sk}	Male	-	0.0383	-	0.0383
Skewness Sk	Female	−0.0312	0	0.0361	0
Standard Error SE_{Sk}	Female	-	0.0549	-	0.0549
Kurtosis K	Male	3.0594	3	3.0295	3
Standard Error SE_K	Male	-	0.0766	-	0.0766
Kurtosis K	Female	3.0222	3	3.1094	3
Standard Error SE_K	Female	-	0.1098	-	0.1098
Bivariate Statistic		r	**Empirical** ρ	**Theory** ρ	**Standard Error** SE_ρ
Pearson Corr. Coeff.	Male	0.4716	0.4689	0.4689	0.0122
	Female	0.5387	0.5335	0.5359	0.0161

It is to be recalled that a random variable X is lognormal if its natural logarithm, symbolized by $Y = \ln X$, is normal. As a matter of standard notation used in this paper, random variables are represented by upper case letters (e.g. X), and realizations of that variable (referred to as variates) are represented by lower case letters (e.g. x). Histograms of the natural logarithms of the ANSUR heights and weights, partitioned by gender, were shown in Part I to be satisfactorily described by PDFs of Gaussian form

$$p_H^{(N)}(y) = \frac{1}{\sqrt{2\pi s_H^2}} \mathrm{e}^{-\frac{(y-m_H)^2}{2s_H^2}} \tag{2}$$

$$p_W^{(N)}(y) = \frac{1}{\sqrt{2\pi s_W^2}} \mathrm{e}^{-\frac{(y-m_W)^2}{2s_W^2}}. \tag{3}$$

A more detailed demonstration of the normality of $\ln H$ and $\ln W$ will be given in Section 6. From relations (2) and (3) follow the parent lognormal PDFs

$$p_H^{(\Lambda)}(x) = \frac{1}{\sqrt{2\pi s_H^2}} \frac{\mathrm{e}^{-\frac{(\ln(x)-m_H)^2}{2s_H^2}}}{x} \tag{4}$$

$$p_W^{(\Lambda)}(x) = \frac{1}{\sqrt{2\pi s_W^2}} \frac{\mathrm{e}^{-\frac{(\ln(x)-m_W)^2}{2s_W^2}}}{x} \tag{5}$$

with location parameters (m_H, m_W) and scale parameters and (s_H, s_W) for height and weight, respectively. Numerical values of these parameters are given in Table 1.

Superscripts N and Λ in the above PDFs signify normal and lognormal distributions, as well as symbolize the associated random variables (RVs)

$$\left.\begin{aligned} X &= \Lambda\left(m, s^2\right) \\ Y &= N\left(m, s^2\right) \end{aligned}\right\} \Rightarrow \begin{cases} Y = \ln X \\ X = \mathrm{e}^Y \end{cases}. \tag{6}$$

Note that parameters $\left(m, s^2\right)$ defining the random variables Y and X are the mean and variance of the *normal* variable Y. *All* statistics of the marginal distribution of X are predictable in terms of the parameters $\left(m, s^2\right)$ of Y [11] [16]. For example, the mean μ_X, variance σ_X^2, skewness Sk_X, and kurtosis K_X of X take the forms [17]

$$\mu_X = \mathrm{e}^{m+\frac{1}{2}s^2}. \tag{7}$$

$$\sigma_X^2 = \mathrm{e}^{2m}\left(\mathrm{e}^{2s^2} - \mathrm{e}^{s^2}\right) \tag{8}$$

$$Sk_X = \left(\mathrm{e}^{s^2} + 2\right)\sqrt{\mathrm{e}^{s^2} - 1} \tag{9}$$

$$K_X = \mathrm{e}^{4s^2} + 2\mathrm{e}^{3s^2} + 3\mathrm{e}^{2s^2} - 3. \tag{10}$$

1.2. Correlation of Height and Weight

When analyzing lognormal variates, it is often strategically easier—indeed necessary—to work with the logarithms of the variates, since these are distributed normally. Figure 2 shows scatter plots (black) of the scaled variates of $Y_W \equiv \ln W$ against scaled variates of $Y_H \equiv \ln H$ for males and females respectively in the ANSUR data set. The scaled variables (U,V)

$$\begin{aligned} U &\equiv \left(\ln H - m_H\right)/s_H \\ V &\equiv \left(\ln W - m_W\right)/s_W \end{aligned} \tag{11}$$

with variates (u,v) are measured with respect to their means and divided by their standard deviations, and are therefore dimensionless quantities distributed as standard normal variables of mean 0 and variance 1 if the variables H, W are lognormal. Figure 2 likewise clearly shows a strong linear correlation of U and V. The slope of the line of regression (dashed red) in each black scatter plot directly yields the corresponding Pearson correlation coefficient r [11] defined by

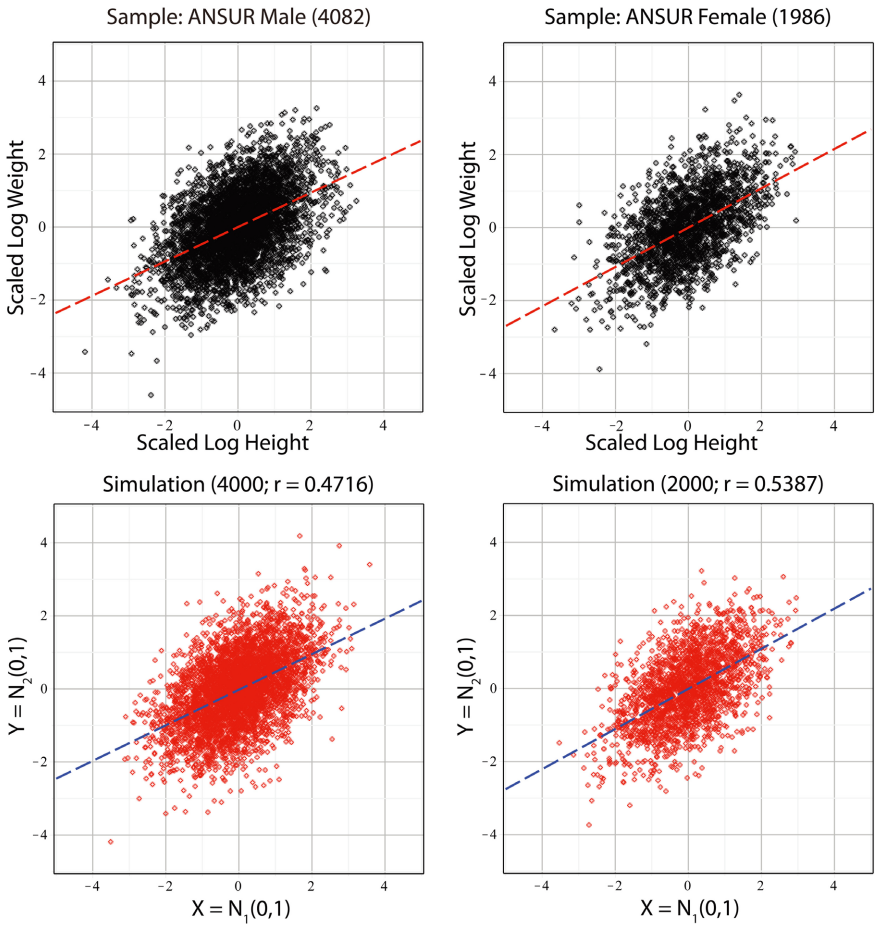


Figure 2. Correlation of scaled log weight and scaled log height for males (top left) and females (top right) of the ANSUR population. Lower panels show corresponding scatter plots created with correlated lognormal random number generators using the same empirical parameters. The patterns display a strong linear correlation. The Pearson correlation coefficient is equal to the slope of the associated lines of regression (dashed).

$$r \equiv \langle UV \rangle = \frac{\langle (Y_H - m_H)(Y_W - m_W) \rangle}{s_H s_W} \equiv \frac{\mathrm{cov}(Y_H, Y_W)}{s_H s_W} \tag{12}$$

where cov signifies covariance, as defined in Equation (12). Angular brackets are used in this paper to indicate expectation values. The scatter plots in red in Figure 2 were obtained by computer simulation using correlated lognormal RNGs, the details of which will be discussed in a later section. Suffice it to say at this point that the computer simulations employed the same distribution parameters that were extracted from the ANSUR height and weight data and are labeled (m_H, m_W, s_H, s_W, r) for both male and female cohorts in Table 1. Lines of regression (dashed blue) to the simulated scatter plots are nearly identical to those of the empirical plots.

Once the correlation coefficient r has been determined empirically from the sample of normal variates (u, v), the correlation coefficient ρ of the parent lognormal variables $X_H \equiv \mathrm{e}^{Y_H} = H$ and $X_W \equiv \mathrm{e}^{Y_W} = W$ can be calculated theoretically from the relation [11]

$$\rho_{thy} = \frac{\mathrm{e}^{r s_H s_W} - 1}{\sqrt{\left(\mathrm{e}^{s_H^2} - 1\right)\left(\mathrm{e}^{s_W^2} - 1\right)}} \tag{13}$$

and compared with the empirical correlation coefficient obtained directly from the data according to

$$\rho_{emp} \equiv \frac{\mathrm{cov}(X_H, X_W)}{\sigma_H \sigma_W} = \frac{\langle (H - \mu_H)(W - \mu_W) \rangle}{\sigma_H \sigma_W} \tag{14}$$

in analogy to Equation (12). The implementation of Equation (14) can be achieved algebraically and geometrically:

1) *Algebraic method*: If h_i and w_i are variates of H and W, such as plotted in Figure 1, where $i = 1, \cdots, n$, then

$$\rho_{emp} = \frac{1}{n \sigma_H \sigma_W} \sum_{i=1}^{n} (h_i - \mu_H)(w_i - \mu_W) \tag{15}$$

in the limit of large n.

2) *Geometric method*: ρ_{emp} is equal to the slope of the line of regression in a scatter plot of the scaled variables $(W - \mu_W)/\sigma_W$ against $(H - \mu_H)/\sigma_H$. To deduce ρ_{emp} from the line of regression in the plot of the unscaled variables (H, W) in Figure 1, one multiplies the slope by the ratio of standard deviations σ_H / σ_W.

Five parameters (m_1, m_2, s_1, s_2, r) are required to specify the PDF of two correlated bivariate normal RVs (Y_1, Y_2)

$$\begin{aligned} p_{Y_1,Y_2}^{(N,N)}(y_1, y_2) &= \frac{1}{2\pi s_1 s_2 \sqrt{1 - r^2}} \mathrm{e}^{-q_Y/2} \\ q_Y &= \frac{1}{1 - r^2}\left[\left(\frac{y_1 - m_1}{s_1}\right)^2 - 2r\left(\frac{y_1 - m_1}{s_1}\right)\left(\frac{y_2 - m_2}{s_2}\right) + \left(\frac{y_2 - m_2}{s_2}\right)^2\right] \end{aligned} \tag{16}$$

from which is derived the PDF of the parent bivariate lognormal RVs (X_1, X_2) [11]

$$p_{X_1,X_2}^{(\Lambda,\Lambda)}(x_1,x_2)=\frac{1}{2\pi s_1 s_2\sqrt{1-r^2}}\frac{\mathrm{e}^{-q_X/2}}{x_1x_2}$$
$$q_X=\frac{1}{1-r^2}\left[\left(\frac{\ln(x_1)-m_1}{s_1}\right)^2-2r\left(\frac{\ln(x_1)-m_1}{s_1}\right)\left(\frac{\ln(x_2)-m_2}{s_2}\right)+\left(\frac{\ln(x_2)-m_2}{s_2}\right)^2\right] \quad (17)$$

Double superscripts N and Λ signify that both variables are normal in PDF (16) and lognormal in PDF (17). The expectation operations in Equation (12) and Equation (14) are performed respectively with PDF (16) and PDF (17).

If, as proposed in this paper, H and W are correlated bivariate lognormal variables, then *all* measurable statistical information concerning adult human height, weight, and their correlations, should be predictable from their joint distribution Equation (17) in terms of the five parameters (2 means, 2 variances, and 1 Pearson correlation) that define a given population. This statement has important implications for the study of obesity and its associated illnesses.

The BMI (1) was introduced by Quetelet in 1835 [18] and has been widely used up to present times by clinicians and epidemiologists as a proxy for obesity under the assumption that it correlates strongly with weight, but is independent of height. This assumption is itself predicated on a by-no-means obvious assumption that human weight in a healthy adult population varies as the *square* of an individual's height. These assumptions will be examined later in this paper both empirically and theoretically. It is to be noted at this point, however, that both assumptions have elicited criticism, e.g. [19] [20] [21], leading to proposals of alternative power-law measures such as the Benn Index [22] and Rohrer's Index [23], non-power law correlations such as [24] [25], and empirical parametric models such as [26], all purporting to determine more satisfactorily than BMI a single optimal relationship between human weight and height.

With regard to the goal of capturing the relationship between height and weight, the following general statistical principles must be emphasized. First, an exact PDF of the bivariate distribution of two correlated random variables provides *all* the statistical information that can be learned about these two correlated variables. And second, there *is no single* optimal mathematical expression—apart from the PDF and its equivalent transformations[1]—that completely captures the statistical relation between two correlated random variables. Rather, the PDF provides a potentially infinite number of mathematical expressions that, *together*, characterize the complete relation between the two variables. From a

[1]These transformed functions are the characteristic function (CF), which is the Fourier transform of the PDF, and the cumulative distribution function (CDF), which is the integral of the PDF from some fixed point to the argument of the PDF. Thus, if $g(x)$ is the CDF, then the PDF $f(x)=\mathrm{d}g(x)/\mathrm{d}x$.

practical standpoint, however, the number of testable expressions that can meaningfully characterize the correlation of two variables is limited by the size of the sample, since the intrinsic uncertainty increases with the order (*i.e.* power) of the variables, and can eventually exceed the mean value for a fixed sample size. These points will be elaborated on in the following sections.

1.3. Organization

The remainder of this paper is organized as follows.

In Section 2 the relation between weight (W) and height (H) is examined by means of the conditional expectation functions of W, given H.

In Section 3 the proposition that human height and weight are correlated lognormal variables is tested by examining generalized correlation functions of data sets (H,W) and $(\ln H,\ln W)$.

In Section 4 the preceding data sets are each examined for nonlinear correlations *beyond* those attributable to the bivariate lognormal distribution by a procedure known as distance correlation.

In Section 5 the marginal distributions of (H,W) and $(\ln H,\ln W)$ are tested against predictions of the univariate lognormal and normal distributions, respectively.

Section 6 examines the implications of the distribution of (H,W) for the body mass index.

Section 7 discusses the computer simulation of correlated lognormal variables.

And last, the results of this comprehensive investigation are summarized and interpreted in Section 8.

2. Conditional Expectation of Weight, Given Height

The conditional expectation $\langle W^p | H \rangle$ of W^p (for $p=1,2,\cdots$) given H, defined by the ratio

$$W_p(h) \equiv \frac{\int_0^\infty w^p\, p_{H,W}^{(\Lambda,\Lambda)}(h,w)\,\mathrm{d}w}{\int_0^\infty p_{H,W}^{(\Lambda,\Lambda)}(h,w)\,\mathrm{d}w}, \tag{18}$$

is a function $W_p(h)$ of the continuous variate of H. Since $W_p(h)$ derives from the joint PDF of W and H, it is more informative than an empirical line of regression such as obtained by the method of least squares or, more generally, the method of maximum likelihood [27].

Calculation of $W_p(h)$ in Equation (18) requires evaluation of two integrals whose kernel is the PDF $p_{H,W}^{(\Lambda,\Lambda)}(h,w)$ given by Equation (17). The integrals can be greatly simplified by the transformation (11) to variables[2] $u=\left(\ln(h)-m_H\right)/s_H$ and $v=\left(\ln(w)-m_W\right)/s_W$, which re-expresses the bivariate normal PDF (16) more simply in the form

[2]Integration variables, in contrast to random variables, will be represented by lower case letters.

$$f_{U,V}(u,v)=\frac{1}{2\pi\sqrt{1-r^2}}\exp\left(-\frac{1}{2\left(1-r^2\right)}\left(u^2-2ruv+v^2\right)\right) \tag{19}$$

and, through the inverse transformation

$$\begin{aligned} h&=\exp\left(m_H+s_H u\right)\\ w&=\exp\left(m_W+s_W v\right)\end{aligned} \tag{20}$$

leads to the conditional expectation

$$W_p(u)\equiv\frac{\int_{-\infty}^{\infty}\left(\exp\left(s_W v+m_W\right)\right)^p f(u,v)\,\mathrm{d}v}{\int_{-\infty}^{\infty}f(u,v)\,\mathrm{d}v} \tag{21}$$

as a function of u. Both integrals in (21) are readily evaluated in closed form., after which replacement of the normal variable u in terms of the lognormal variable h yields the general relation

$$W_p(h)=h^{prs_W/s_H}\exp\left(pm_W+\frac{1}{2}p^2s_W^2\left(1-r^2\right)-prm_Hs_W/s_H\right). \tag{22}$$

The conditional expectation of the p^{th} power of W is thus seen to have a power-law dependence on H with exponent prs_W/s_H. The lowest two orders p = 1, 2 are of primary interest

$$W_1(h)=h^{\frac{rs_W}{s_H}}\exp\left(m_W+\frac{1}{2}s_W^2\left(1-r^2\right)-\frac{rm_Hs_W}{s_H}\right) \tag{23}$$

$$W_2(h)=h^{\frac{2rs_W}{s_H}}\exp\left(2m_W+2s_W^2\left(1-r^2\right)-\frac{2rm_Hs_W}{s_H}\right) \tag{24}$$

and yield the conditional variance and standard deviation

$$\begin{aligned}\mathrm{var}(W\mid H)&=W_2(h)-\left(W_1(h)\right)^2\\&=h^{\frac{2rs_W}{s_H}}\mathrm{e}^{\left(2m_W-\frac{2rm_Hs_W}{s_H}\right)}\left(\mathrm{e}^{2s_W^2\left(1-r^2\right)}-\mathrm{e}^{s_W^2\left(1-r^2\right)}\right)\end{aligned} \tag{25}$$

$$\begin{aligned}\sigma_W(h)&\equiv\sigma(W\mid H)=\sqrt{\mathrm{var}(W\mid H)}\\&=h^{\frac{rs_W}{s_H}}\mathrm{e}^{\left(m_W-\frac{rm_Hs_W}{s_H}\right)}\left(\mathrm{e}^{2s_W^2\left(1-r^2\right)}-\mathrm{e}^{s_W^2\left(1-r^2\right)}\right)^{\frac{1}{2}}\end{aligned} \tag{26}$$

Substituting in Equations (23) and (26) the bivariant lognormal parameters for each gender cohort of the ANSUR sample listed in Table 1 leads to the expressions

$$\begin{aligned}\left(W_1(h)\pm\sigma_W(h)\right)_{\mathrm{M}}&=\left(27.7071\pm4.0621\right)h^{1.9987}\\\left(W_1(h)\pm\sigma_W(h)\right)_{\mathrm{F}}&=\left(23.2156\pm3.1523\right)h^{2.1923}\end{aligned} \tag{27}$$

where subscripts M and F signify male and female, respectively.

Plots of the conditional expectations $W_1(h)$ in Equation (27) comprise the solid blue curves superposed on the scatter plots in Figure 1. Although $W_1(h)$

is a power law and the line of regression (dashed red) is linear, the two curves are virtually indistinguishable over the densest part of the plots. However, it is important to bear in mind the conceptual difference between the two curves: the line of regression is merely a fit to data, whereas the mathematical relation (22), the specific exponent prs_W/s_H, and the numerical values in expressions (27) are *predictions* drawn from the lognormal PDF. Figure 3 graphically displays the full information content of relation (27) by displaying the regions of ±1 standard deviation about the means for the two cohorts.

The numerical values of the exponents in relations (27) bear out the fundamental assumption underlying the use of BMI that weight is a quadratic function of height in a healthy adult population. Nevertheless, for a different set of values of the parameters (s_H, s_W, r), such as may characterize a demographic different from the one represented by the ANSUR population, the lognormal predicted exponent could be different.

3. Tests of Correlation Functions of Height (H) and Weight (W)

Correlation functions of order (p, q), defined as follows

$$C_{p,q}(s_H, s_W, r) \equiv \left\langle \left(\frac{H-\mu_H}{\sigma_H} \right)^p \left(\frac{W-\mu_W}{\sigma_W} \right)^q \right\rangle \tag{28}$$

$$R_{p,q}(r) \equiv \left\langle \left(\frac{\ln(H)-m_H}{s_H} \right)^p \left(\frac{\ln(W)-m_W}{s_W} \right)^q \right\rangle, \tag{29}$$

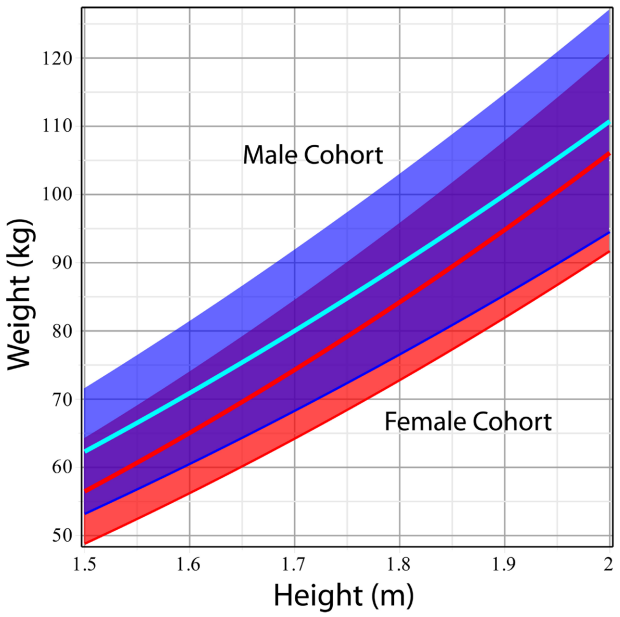


Figure 3. Plots of the conditional expectation $W_1(h)$ for male (solid blue line) and female (solid red line) of the ANSUR populations centered on regions (blue for males, red for females) of ±1 standard deviation $\sigma_W(h)$. The region of overlap appears purple.

generalize the standard covariance ($p=q=1$) expressed in relations (12) and (14). Expectations (28) and (29) are to be implemented with the bivariate lognormal PDF (17) where indices (p, q) independently take on integer values (1, 2, …).

As a matter of notation and terminology, evaluation of functions $C_{p,q}$ and $R_{p,q}$ by substitution of empirical parameters for the arguments yield the numerical correlation coefficients $c_{p,q}$, $r_{p,q}$, where $r_{1,1}\equiv r$ and $c_{1,1}\equiv\rho$ as conventionally defined. If the proposition that (H,W) are bivariate lognormal variables is valid, then $c_{p,q}$ and $r_{p,q}$ should be predictable from the arguments shown in Equations (28) and (29) and the parameters (s_H,s_W,r) listed by gender in Table 1. It is to be recalled that the first two parameters (standard deviations) were obtained empirically from the variates of the marginal distributions $\ln H$ and $\ln W$ of the ANSUR populations, whereas the third parameter (Pearson correlation coefficient) was obtained empirically from the joint distribution of $\ln H$ and $\ln W$, such as exhibited in Figure 2. The associated sets of means (m_H,m_W), which are also listed by gender in Table 1, drop out of relations (28) and (29) by virtue of their expressions as ratios. Further details are given in Part I [11]. To facilitate reading the tables to follow, indices of the correlation functions and coefficients will be expressed as arguments, e.g. $R(p,q)$ and $r(p,q)$ in the tables.

Calculation of the correlation functions $C_{p,q}$ and $R_{p,q}$ requires evaluation of the expectation values

$$C_{p,q}=\int_0^\infty\int_0^\infty\left(\frac{h-\mu_H}{\sigma_H}\right)^p\left(\frac{w-\mu_W}{\sigma_W}\right)^q p_{(H,W)}^{(\Lambda,\Lambda)}(h,w)\,\mathrm{d}h\mathrm{d}w \tag{30}$$

$$R_{p,q}=\int_0^\infty\int_0^\infty\left(\frac{\ln(h)-m_H}{s_H}\right)^p\left(\frac{\ln(w)-m_W}{s_W}\right)^q p_{(H,W)}^{(\Lambda,\Lambda)}(h,w)\,\mathrm{d}h\mathrm{d}w. \tag{31}$$

As in the previous section, the two integrals can be greatly simplified by the transformation (11) to variables $u=\left(\ln(h)-m_H\right)/s_H$ and $v=\left(\ln(w)-m_W\right)/s_W$, which results in the bivariate normal PDF (19). Substitution for the means (μ_H,μ_W) and standard deviations (σ_H,σ_W) by use of Equations (7) and (8) then leads to the operational expressions

$$C_{p,q}=\left(\mathrm{e}^{s_H^2}-1\right)^{-p/2}\left(\mathrm{e}^{s_W^2}-1\right)^{-q/2}\int_{-\infty}^\infty\int_{-\infty}^\infty\left(\mathrm{e}^{s_Hu-\frac{1}{2}s_H^2}-1\right)^p\left(\mathrm{e}^{s_Wv-\frac{1}{2}s_W^2}-1\right)^q f_{U,V}(u,v)\,\mathrm{d}u\mathrm{d}v \tag{32}$$

and

$$R_{p,q}=\int_{-\infty}^\infty\int_{-\infty}^\infty u^pv^q f_{U,V}(u,v)\,\mathrm{d}u\mathrm{d}v. \tag{33}$$

In the symmetric case ($p=q$), the two correlation functions will simply be designated C_p and R_p.

The integral in relation (33) displays a number of symmetries: 1) $R_{p,q}=R_{q,p}$, 2) $R_{p,q}=0$ if $p+q$ is odd, otherwise 3) if $p+q$ is even, then $R_{p,q}$ is of

order r^d where d is the smaller of p and q. Symmetries (2) and (3) do not necessarily hold for $C_{p,q}$. As a matter of terminology in the sections to follow, $R_{p,q}$ and $C_{p,q}$ will be referred to as "asymmetric odd" if $p+q$ is odd and "asymmetric even" if $p+q$ is even.

3.1. Calculation and Measurement of Correlation Functions $R_{p,q}$

Because the correlation of U and V is determined entirely by the parameter r in the probability density $f(u,v)$, it is useful to examine the structure of $f(u,v)$ graphically. The left panels of Figure 4 display a sequence of theoretical 3-dimensional density plots for correlation coefficients r = 0.0, 0.5, and 0.95, which span nearly the entire positive range of r. The maximum r = 0.95 was chosen, rather than $r=1$, because the latter value is simply a straight line coincident with the diagonal axis. The right panels view the underside of the density plots, or, equivalently, the projection of the plots onto the (u,v) plane. In the top panels, vectors U and V are independent ($r=0$), and $f_{U,V}(u,v)$ factors into a product $f_U(u)f_V(v)$. Since the exponential in $f_{U,V}(u,v)$ for $r=0$ is the sum u^2+v^2, a transformation of variables converts that sum into the square of a radial variable, which accounts for the circular symmetry of the top right panel. As the correlation coefficient increases toward +1, the density function profile becomes increasingly linear along the diagonal for which $uv>0$, as clearly shown in the bottom panels. For $r\to -1$, the profile (not shown) would approach linearity along the diagonal for which $uv<0$.

Whereas the left side of Figure 4 shows the actual probability density profiles, the images on the right side show smooth shapes to which scatter plots of a sample of discrete paired variates (u_i,v_i), $i=1,\cdots,n$, approach as $n\to\infty$. This is borne out by Figure 5, which shows simulated plots of n = 2000 pairs of correlated standard normal variates of increasing correlation coefficient r. In each panel, the quantity "rho" designates the correlation parameter supplied to the RNGs; the quantity "slope" is the slope of the line of regression (dashed blue), which equals the actual correlation coefficient of the simulated sample. The two numbers in each scatter plot are close, but not identical because the scatter plot comprises a *finite* random sample. The connection between the PDF $f_{U,V}(u,v)$ and the scatterplot of (U,V) for each value of r is particularly clear when one compares the right side panels of Figure 4 to the corresponding plots of Figure 5. The orientations of the two sets of figures may be different, but the distributions are invariant to orientation.

Correlation function $R_{p,q}(r)$ expressed in Equation (33) can be evaluated in closed form, and depends only on the Pearson coefficient r. Table 2 lists the first six orders of the symmetric correlation functions and the most pertinent of the asymmetric correlation functions. It is seen that beyond the basic covariance (12), the higher-order symmetric correlations are increasingly nonlinear in r. Asymmetric correlation functions of the form $R_{p,1}$ in Table 2, where p is the exponent of the scaled variable for $\ln H$, address the issue (referred to in the

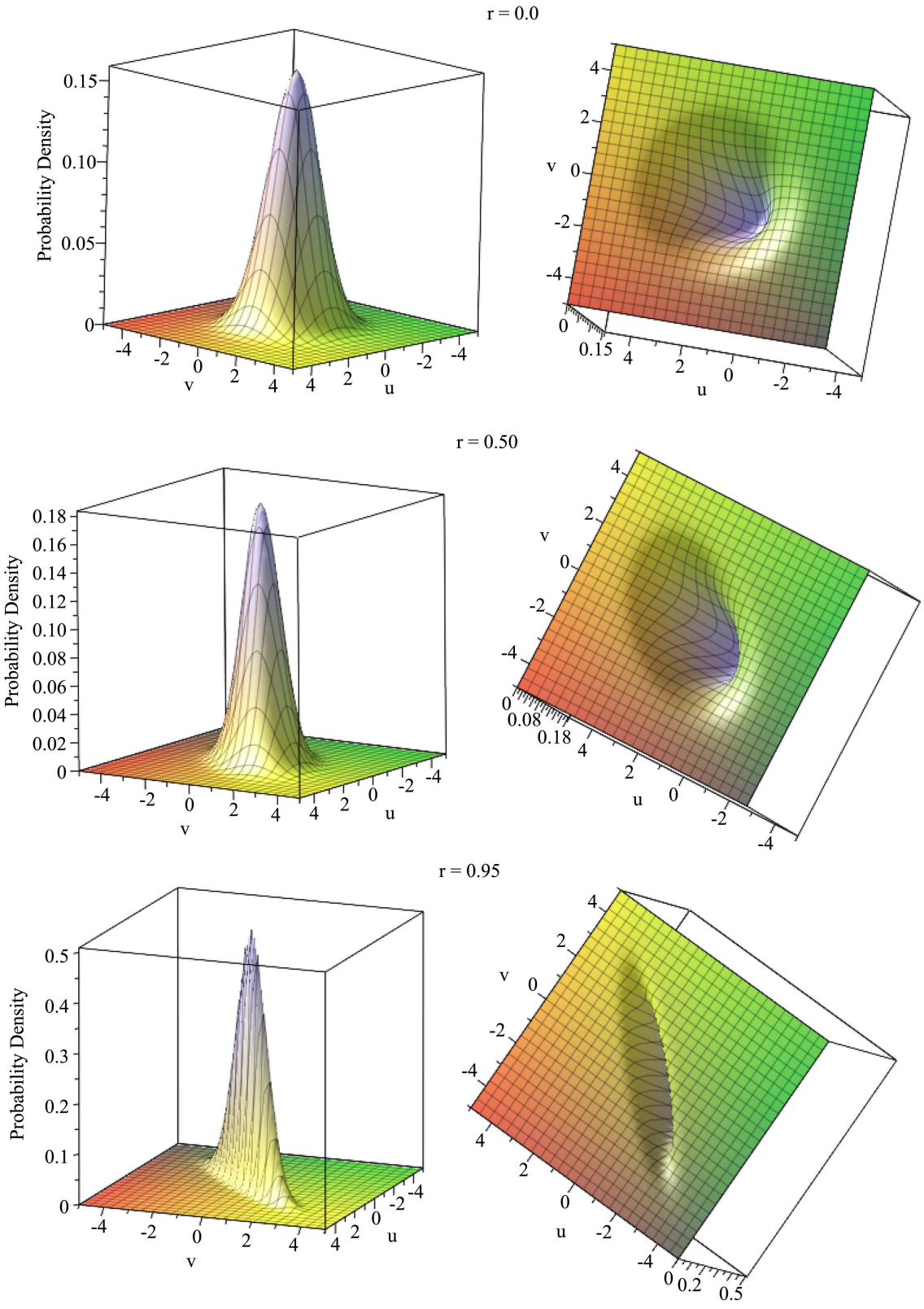


Figure 4. Left Panels: Probability density profiles $f(u,v)$ of correlated standard normal variates (u,v) for correlation coefficients $r = 0$ (top), 0.5 (center), 0.95 (bottom). Right Panels: Profiles as viewed from the underside of the associated density plots. The patterns are the smooth shapes of scatter plots of discrete samples in the limit of infinite sample size.

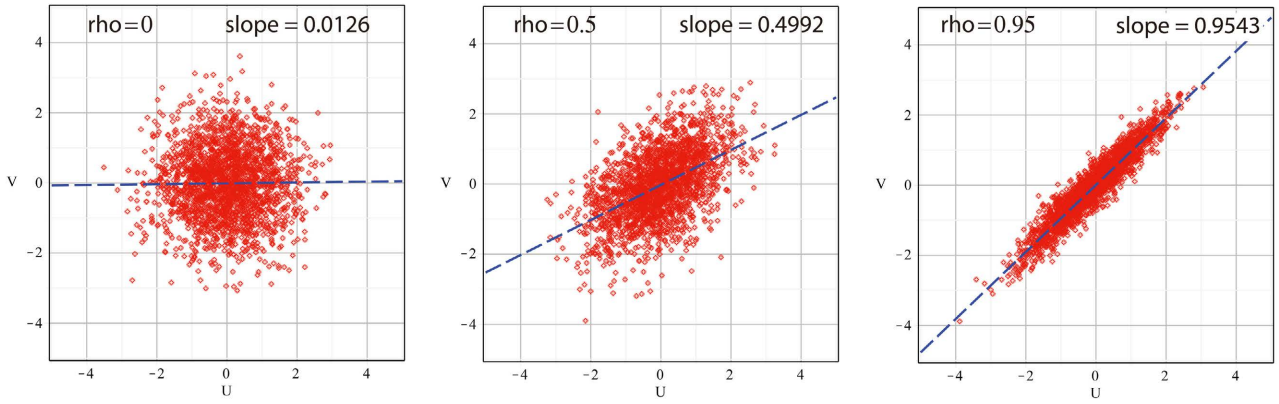


Figure 5. Simulated scatter plots of 2000 correlated pairs of standard normal variates with correlation coefficients increasing from 0 to nearly 1. In each plot, "rho" is the numerical value of the correlation parameter supplied to the RNGs, and "slope" (of the line of regression) is the actual correlation coefficient produced by the simulation.

Table 2. Correlation functions $R(p, q)$ of powers of standard normal variables: U^p and V^q.

Symmetric	Expectation Value	Variance
$R(1, 1)$	r	$1+r^2$
$R(2, 2)$	$1+2r^2$	$4\left(2+17r^2+5r^4\right)$
$R(3, 3)$	$9r+6r^3$	$9\left(25+441r^2+588r^4+76r^6\right)$
$R(4, 4)$	$9+72r^2+24r^4$	$12^2\left(76+2441r^2+7311r^4+3896r^6+276r^8\right)$
$R(5, 5)$	$225r+600r^3+120r^5$	$15^2\left(3969+198225r^2+1057200r^4+1268240r^6+362240r^8+16064r^{10}\right)$
$R(6, 6)$	$225+4050r^2+5400r^4+720r^6$	$90^2\left(13334+960273r^2+8001825r^4+17070080r^6+10972800r^8+1950528r^{10}+59072r^{12}\right)$
Asymmetric		
$R(2, 1)$	0	$3+12r^2$
$R(3, 1)$	$3r$	$15+81r^2$
$R(4, 1)$	0	$105+840r^2$
$R(5, 1)$	$15r$	$3^2\left(105+1025r^2\right)$
$R(6, 1)$	0	$3^2\left(1155+13860r^2\right)$
$R(3, 2)$	0	$3^2\left(5+60r^2+40r^4\right)$
$R(4, 2)$	$3+12r^2$	$3^2\left(34+552r^2+544r^4\right)$
$R(4, 3)$	0	$3^2\left(175+4200r^2+8400r^4+2240r^6\right)$
$R(5, 3)$	$45r+60r^3$	$15^2\left(63+1881r^2+5016r^4+2000r^6\right)$
$R(6, 4)$	$45+540r^2+360r^4$	$45^2\left(538+25848r^2+129200r^4+137792r^6+29504r^8\right)$

Introduction) concerning how weight correlates with powers of height. Also listed are asymmetric correlation functions characterizing how powers of weight correlate with powers of height.

Although the Pearson correlation coefficient r is the expectation value of a composite random variable UV, it is regarded here as a nondistributed quantity since the values of r for male and female cohorts used throughout this paper are fixed parameters extracted from the ANSUR data. The same is true of the means (m_H, m_W) and variances (s_H^2, s_W^2) for the male and female cohorts. The variance of the correlation coefficient $r_{p,q}$ therefore characterizes only the variation of the product $U^p V^q$ in the defining integral and can be expressed by [28]

$$\mathrm{var}\left(r_{p,q}\right) = r_{2p,2q} - r_{p,q}^2 . \tag{34}$$

It then follows that the standard error (se) of $r_{p,q}$ takes the form

$$se\left(r_{p,q}\right) = \sqrt{\frac{\mathrm{var}\left(r_{p,q}\right)}{n}} = \sqrt{\frac{r_{2p,2q} - r_{p,q}^2}{n}} \tag{35}$$

where n is the sample size. By the same reasoning, the standard error of the correlation coefficient $c_{p,q}$ is

$$se\left(c_{p,q}\right) = \sqrt{\frac{\mathrm{var}\left(c_{p,q}\right)}{n}} = \sqrt{\frac{c_{2p,2q} - c_{p,q}^2}{n}} \tag{36}$$

since $c_{p,q}$ likewise depends on the fixed ANSUR parameters. In general, however, the distribution function and moments of even the lowest correlation coefficient $r_{1,1}$ are difficult to obtain in closed form [29], and, at the time of writing, the author knows of no calculation in the literature of the exact closed-form distribution functions and moments of higher-order correlation coefficients of two normal or two lognormal variables.

Theoretical and empirical ANSUR correlation coefficients $r_{p,q}$ are displayed in **Table 3** for symmetric indices up to $p = 6$ and for a selection of asymmetric indices ranging from (2, 1) to (6, 4). For other correlation orders, the corresponding standard errors were too large relative to the means for a comparison of theory and experiment to be meaningful. Examination of **Table 3** shows striking agreement between lognormal theory and the ANSUR data.

For all but two correlation coefficients of the female cohort listed in the table, the magnitude of the difference between theoretical (thy) and empirical (emp) coefficients did not exceed 1 standard error. In other words, assuming, as justified by the Central Limit Theorem [30], that the relative error $z = \left(r_{p,q}\big|_{emp} - r_{p,q}\big|_{thy}\right) \Big/ se\left(r\right)$ follows a normal distribution, then rejection of the hypothesis $r_{p,q}\big|_{emp} = r_{p,q}\big|_{thy}$ at the conventional 5% threshold would require $|z| \geq 1.65$ [31], which was *not* the case for any of the correlation coefficients of the female cohort. The relative errors of the two exceptional coefficients $r_{3,2}$ and $r_{4,3}$ were approximately 1.074 and 1.243, respectively.

Table 3. Correlation coefficients $r(p, q)$ of (Scaled $\ln H)^p$ and (Scaled $\ln W)^q$.

Correlation of Order (p, q)	ANSUR Male (Nm = 4082)		ANSUR Female (Nf = 1986)	
Symmetric	**Theory**	**Empirical**	**Theory**	**Empirical**
$r(1, 1)$	0.4716 ± 0.0173	0.4716	0.5387 ± 0.0254	0.5387
$r(2, 2)$	1.4448 ± 0.0769	1.5217	1.5804 ± 0.1208	1.6517
$r(3, 3)$	4.8736 ± 0.5808	5.7232	5.7832 ± 0.9517	6.3744
$r(4, 4)$	26.1998 ± 6.0103	36.9508	31.9151 ± 10.2693	34.9686
$r(5, 5)$	171.8363 ± 79.6825	317.1787	220.4467 ± 141.8672	228.2331
$r(6, 6)$	1400.7283 ± 1289.8011	3411.6961	1872.6315 ± 2392.8292	1705.9690
Asymmetric (even $p + q$)				
$r(3, 1)$	1.4148 ± 0.0899	1.4668	1.6161 ± 0.1384	1.6800
$r(5, 1)$	7.0739 ± 0.8568	7.8256	8.0804 ± 1.3408	8.3927
$r(7, 1)$	49.5172 ± 11.6439	60.4761	56.5632 ± 18.3731	56.2658
$r(4, 2)$	5.6688 ± 0.6364	6.4405	6.4823 ± 1.0323	6.4405
$r(5, 3)$	27.5146 ± 6.4358	37.7596	33.6210 ± 10.9103	37.7596
r (6, 4)	182.9017 ± 148.8452	330.3891	232.0216 ± 148.8452	242.3626
Asymmetric (odd $p + q$)				
$r(2, 1)$	0 ± 0.0373	-0.0435	0 ± 0.0568	-0.0032
$r(4, 1)$	0 ± 0.2674	-0.5132	0 ± 0.4162	-0.3299
$r(6, 1)$	0 ± 3.0566	−6.8819	0 ± 4.8072	−4.7772
$r(3, 2)$	0 ± 0.2117	−0.3896	0 ± 0.3387	−0.3637
$r(4, 3)$	0 ± 1.8482	−5.1806	0 ± 3.0866	−3.8362
$r(5, 4)$	0 ± 21.6387	−64.2309	0 ± 37.6773	−37.6538

Relative errors of the male cohort were overall larger than those of the female cohort, although most did not exceed the 5% threshhold value of 1.65. Exceptions occurred primarily among the higher orders of the asymmetric odd coefficients whose theoretical means were zero and standard errors large. Under such circumstances, large deviations are to be expected and would require a larger sample size for resolution. Moreover, expressions (35) and (36) give *lower* limits for the standard errors since, in accordance with stated assumptions, they do not take account of the variation in lognormal parameters. It is therefore likely that a more exact estimate of the relative errors would be lower than those listed in Table 3.

In summary, to appreciate how extensive and close is the agreement of the theoretical and empirical correlations displayed in Table 3, one must bear in

mind the following context. Theoretical predictions of $r_{p,q}$ were obtained from the *bivariate normal* distribution of $\ln H$ and $\ln W$; empirical values of $r_{p,q}$ were obtained from the *natural logarithms* of the raw ANSUR sample. If adult heights and weights were not distributed lognormally, then the comprehensive correspondence *by pure chance*, especially in the female cohort, of these two sets of numbers would be extremely improbable. For example, if either or both of the attributes of height and weight were themselves normally distributed, as had long been assumed, the logarithm of the variates would depart significantly from a normal distribution, as was demonstrated in Part I [11].

Nevertheless, the results in **Table 3** raise a curious question. Why does the female cohort appear to bear out the predictions of lognormal theory more closely than the male cohort despite the fact that the number of men sampled is about twice that of women? As discussed in Part I [11], the Anthropometric Survey of U.S. Army Personnel was undertaken to obtain data representative of the "Total Army" [15] with regard to making accurate decisions concerning clothing, protective equipment, workspaces, and other size-dependent, work-related matters. The survey measured more than 90 human attributes directly and compiled data demographically in terms of race, ethnicity, gender, age, and geographic location. For the analyses in this paper and in Part I, the data were partitioned by gender only. Therefore, both the male and female cohorts can be regarded as diverse samples of fundamentally healthy adults. However, since there are considerably fewer women in the U.S. Army than men, it is conceivable that, irrespective of other demographic characteristics, the women who joined the U. S. Army and took part in the ANSUR sample formed a more homogeneous group in regard to body type and physical fitness than the men. Such an explanation would seem likely, since there is no biophysical basis to believe that male height and weight would be statistically distributed by a probability function of different mathematical form than female height and weight.

3.2. Density Plots Associated with Correlation Functions $R_{p,q}$

Whereas the correlation coefficients $r_{p,q}$ are single numbers quantifying the correlation of U^p and V^q (*i.e.* powers of the scaled variables of $\ln H$ and $\ln W$) for a specified sample, the actual scatter plots of the two sets of variates yield a more comprehensive visual perspective of their correlation. **Figure 6** shows such plots for symmetric correlation coefficients of orders $p = 2, 3, 4, 5$ of the ANSUR male cohort (sample size 4082). The patterns for the female cohort are similar, although less dense (sample size 1986) and not shown.

The correlations expressed in **Figure 6** are highly nonlinear in two ways. Geometrically, the overall patterns of order $p > 1$ do not show a well-defined linear variation such as seen in **Figure 2** for $p = 1$. And algebraically, the associated theoretical correlation function is of order r^p in the Pearson correlation parameter, as summarized in **Table 2**. The four plots in **Figure 6** illustrate the characteristic property that, with increasing order p, the density of points clusters

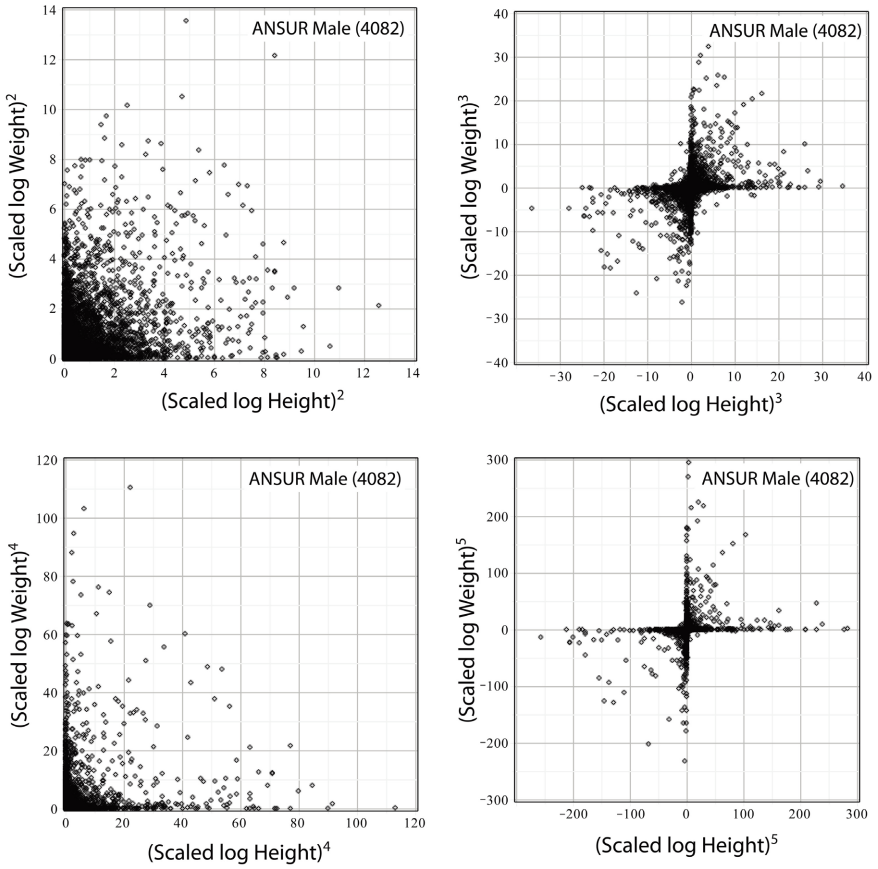


Figure 6. Empirical scatter plots of V^p against U^p for $p = 2$ (top left), 3 (top right), 4 (bottom left), 5 (bottom right), compiled from the ANSUR data. Patterns show a highly nonlinear correlation of weight and height. U is the scaled variable for $\ln H$; V is the scaled variable for $\ln W$.

more tightly about the coordinate axes. Fluctuations for even p extend primarily into the first quadrant (since both variates are positive). Fluctuations for odd p extend primarily into the first and third quadrants for $r > 0$ (and into the second and fourth quadrants for $r < 0$; not shown).

The empirical patterns and properties in **Figure 6** are reproduced nearly identically (apart from random fluctuations) in the computer simulated patterns shown in **Figure 7**. The simulations were created by means of correlated lognormal RNGs using the ANSUR lognormal parameters in **Table 1**. The reproduction of empirical correlation scatter plots (and correlation coefficients) by computer simulation extends as well to asymmetric even and odd correlation functions, as displayed in **Figure 8** for the pairs of variables $\left(U^2,V\right)$, $\left(U^3,V\right)$, and $\left(U^3,V^2\right)$. Plots in black are empirical; those in red are simulated for a population of corresponding size. Altogether, these results support the conclusion that the nonlinear correlations of height and weight stem exclusively from the properties of the lognormal distribution function and depend on no correlation parameters other than the Pearson coefficient r.

Probability density functions for the correlated powers $\left(U^p,V^q\right)$, which yield

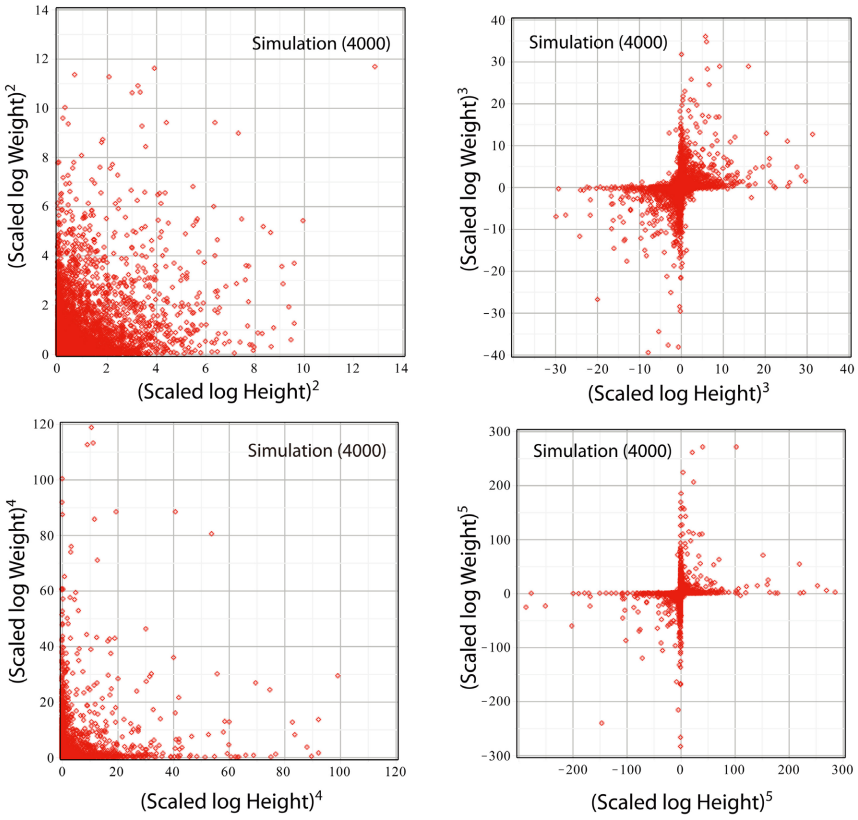


Figure 7. Simulated scatter plots of V^p against U^p for $p = 2$ (top left), 3 (top right), 4 (bottom left), 5 (bottom right), obtained from correlated lognormal RNGs using the lognormal parameters in Table 1. The shapes of the patterns and extent of fluctuations closely resemble the corresponding empirical plots in Figure 6.

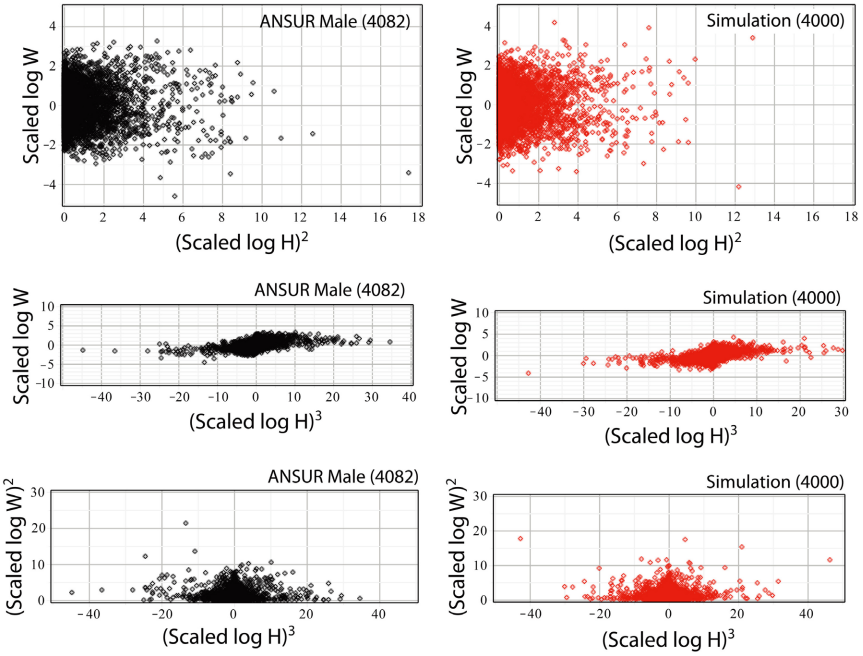


Figure 8. Empirical (black) and simulated (red) scatter plots of V^q against U^p for the asymmetric orders $(p, q) = (2, 1)$ (top), (3, 1) (middle), (3, 2) (bottom).

the patterns approached by scatter plots in Figures 6-8 in the limit of infinite sample size, can be constructed by means of the Dirac delta function as follows

$$f_{X,Y}^{(p,q)}(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{U,V}(u,v)\delta\left(u^p-x\right)\delta\left(v^q-y\right)\mathrm{d}u\mathrm{d}v\,. \tag{37}$$

The subscripts (X,Y) represent the random variables whose lower-case variates are respectively $x=u^p$, $y=v^q$. Powers of standard normal variables like U, V do not, in general, follow known, named distributions to which variables X and Y can be assigned [32]. A brief recapitulation of the properties and identities of the Dirac delta function is given in Part I [11] and in mathematical physics books [33]. The integral (37) can be evaluated in closed form, but gives rise to long, cumbersome expressions for $p>2$, which will not be reproduced here.

Plots of probability density (37) for the symmetric cases $p=2$ and $p=3$ are shown respectively in Figure 9 and Figure 10. Left-side panels show the density

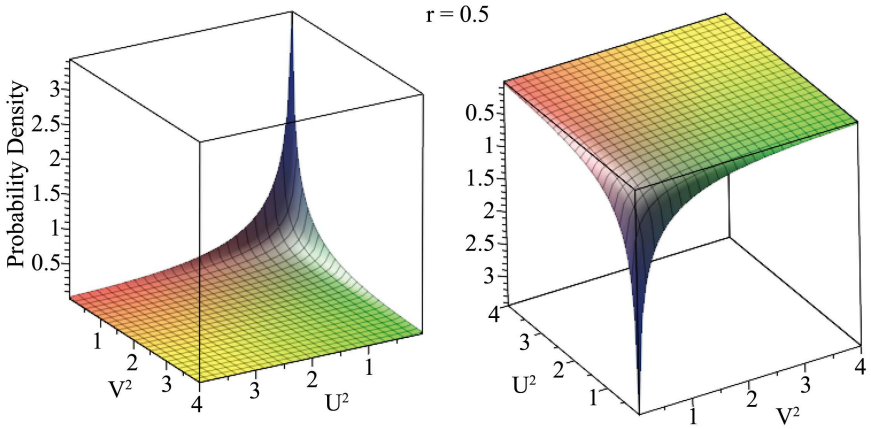


Figure 9. Left panel: Plot of the probability density $f_{X,Y}^{(2)}(u^2,v^2)$ for $r=0.5$. Right panel: View from the underside highlights the profile to which the scatter plot of V^2 against U^2 in the top left of Figure 6 approches in the limit of infinite sample size.

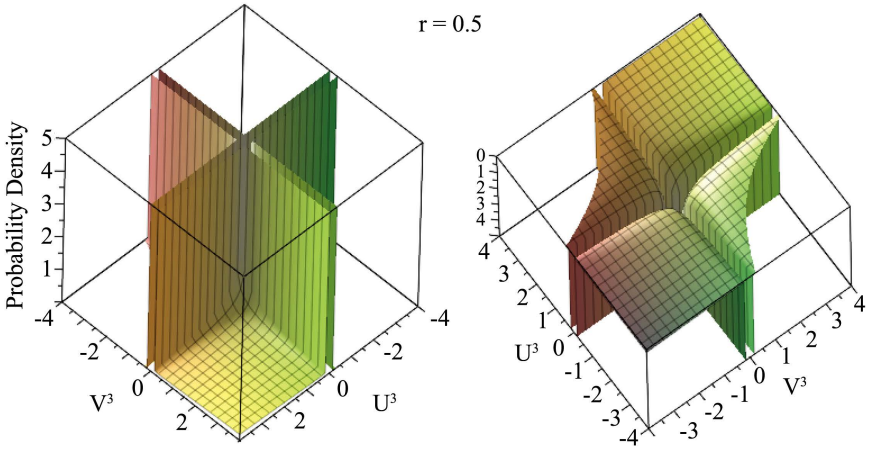


Figure 10. Left panel: Plot of the probability density $f_{X,Y}^{(3)}(u^3,v^3)$ for $r=0.5$. Right panel: View from the underside shows the profile approached by the scatter plot of V^3 against U^3 in the top right of Figure 6 in the limit of infinite sample size.

patterns from above the $\left(u^p, v^p\right)$ plane; right-side panels give complementary images from below. The function $f_{X,Y}^{(2)}\left(u^2, v^2\right)$ in Figure 9 shows a concentration of probability along the vertical axis with density decreasing with distance into the first quadrant, as shown empirically in Figure 7 (top left). Function $f_{X,Y}^{(3)}\left(u^3, v^3\right)$ in Figure 10 (left side) sharply delineates the "cross" of probability density along the coordinate axes, whereas the projection of the pattern onto the $\left(u^3, v^3\right)$ plane (right side) captures the point distribution in the corresponding plot of Figure 7 (top right).

3.3. Calculation and Measurement of Correlation Functions $C_{p,q}$

$C_{p,q}\left(s_H, s_W, r\right)$ in Equation (32) expresses the correlation of height and weight directly, rather than of their logarithms. The integral in Equation (32) can be evaluated in closed form of which the lowest orders pertinent to this paper are given in Table 4 for two arbitrary, but correlated, lognormal variables X_1 and X_2 with bivariate parameter set $\left(m_1, m_2, s_1, s_2, r\right)$. Beyond the symmetric order $p = 4$ and asymmetric order $(p,q) = (4,2)$ expressions for the functions become overly long and not especially informative. Some points to note: 1) symmetric functions $C_p\left(s_H, s_W, r\right)$ are invariant under the interchange of parameters s_H and s_W; 2) asymmetric odd functions $C_{p,q}$ do not identically vanish, as do asymmetric odd $R_{p,q}$; 3) none of the functions $C_{p,q}$ vanishes for $r = 0$, as do the functions $R_{p,q}$ (except those of the form $R(2m, 2n)$ where m, n are integers > 0).

Substitution of the ANSUR parameters of Table 1 for male and female cohorts into the functions of Table 4 and variance of Equation (36) yield the corresponding empirical correlation coefficients summarized in Table 5. Agreement of the empirical values with lognormal theory is again excellent, apart from the highest orders where the standard errors are large relative to the means and signify that a larger sample size is required.

4. Test for Nonlinear Correlations by the Method of Distance Correlation

If adult human height and weight are bivariate lognormal variables, then all measures of their correlation must be calculable from the PDF (17). Evidence for the proposition of bivariate lognormality has been supported up to this point by the agreement of measured and lognormally predicted correlation coefficients and comparison of empirical and lognormally simulated probability density plots. The question remains, however, as to whether there may be nonlinear correlations beyond those intrinsic to the bivariate lognormal distribution. Correlation of distances [12], provides a sensitive method for testing the independence of random vectors.

As initially presented by its developers, the term distance covariance (dCov) of two random vectors X and Y, defined by

$$V(X,Y) \equiv \left\| g_{X,Y}(t,s) - g_X(t)\, g_Y(s) \right\|, \tag{38}$$

Table 4. Correlation functions $C(p, q)$ of Lognormal variables X_1^p and X_2^q (Notation: $E(x) \equiv \exp(x)$).

Moments	**Bivariate parameters** (m_1, m_2, s_1, s_2, r)	
Means	$\mu(X_1) = E\left(m_1 + \frac{1}{2}s_1^2\right)$	$\mu(X_2) = E\left(m_2 + \frac{1}{2}s_2^2\right)$
Variances	$\sigma^2(X_1) = E(2m_1)\left(E\left(2s_1^2\right) - E\left(s_1^2\right)\right)$	$\sigma^2(X_2) = E(2m_2)\left(E\left(2s_2^2\right) - E\left(s_2^2\right)\right)$
Symmetric	**Expectation Value**	
$C(1, 1)$	$\left(\left(e^{s_1^2} - 1\right)\left(e^{s_2^2} - 1\right)\right)^{-1/2}\left(e^{r s_1 s_2} - 1\right)$	
$C(2, 2)$	$\left(\left(e^{s_1^2} - 1\right)\left(e^{s_2^2} - 1\right)\right)^{-1}\left[E\left(s_1^2 + s_2^2 + 4r s_1 s_2\right) - 2E\left(s_1^2 + 2r s_1 s_2\right)\right.$ $\left. - 2E\left(s_2^2 + 2r s_1 s_2\right) + 4E\left(r s_1 s_2\right) + E\left(s_1^2\right) + E\left(s_2^2\right) - 3\right]$	
$C(3, 3)$	$\left(\left(e^{s_1^2} - 1\right)\left(e^{s_2^2} - 1\right)\right)^{-3/2}\left[E\left(3s_1^2 + 3s_2^2 + 9r s_1 s_2\right) - 3E\left(3s_1^2 + s_2^2 + 6r s_1 s_2\right)\right.$ $- 3E\left(s_1^2 + 3s_2^2 + 6r s_1 s_2\right) + 3E\left(3s_1^2 + 3r s_1 s_2\right) + 3E\left(3s_2^2 + 3r s_1 s_2\right)$ $+ 9E\left(s_1^2 + s_2^2 + 4r s_1 s_2\right) - 9E\left(s_1^2 + 2r s_1 s_2\right) - 9E\left(s_2^2 + 2r s_1 s_2\right)$ $\left. + 9E\left(r s_1 s_2\right) - E\left(3s_1^2\right) - E\left(3s_2^2\right) + 3E\left(s_1^2\right) + 3E\left(s_2^2\right) - 5\right]$	
$C(4, 4)$	$\left(\left(e^{s_1^2} - 1\right)\left(e^{s_2^2} - 1\right)\right)^{-3/2}\left[E\left(6s_1^2 + 6s_2^2 + 16r s_1 s_2\right) + 16E\left(3s_1^2 + 3s_2^2 + 9r s_1 s_2\right)\right.$ $+ 36E\left(s_1^2 + s_2^2 + 4r s_1 s_2\right) + 6E\left(s_1^2 + 6s_2^2 + 8r s_1 s_2\right) + 6E\left(6s_1^2 + s_2^2 + 8r s_1 s_2\right)$ $- 4E\left(6s_1^2 + 3s_2^2 + 12r s_1 s_2\right) - 4E\left(3s_1^2 + 6s_2^2 + 12r s_1 s_2\right) - 24E\left(3s_1^2 + s_2^2 + 6r s_1 s_2\right)$ $- 24E\left(s_1^2 + 3s_2^2 + 6r s_1 s_2\right) - 4E\left(6s_1^2 + 4r s_1 s_2\right) - 4E\left(6s_2^2 + 4r s_1 s_2\right)$ $+ 16E\left(3s_1^2 + 3r s_1 s_2\right) + 16E\left(3s_2^2 + 3r s_1 s_2\right) - 24E\left(s_1^2 + 2r s_1 s_2\right) - 24E\left(s_2^2 + 2r s_1 s_2\right)$ $\left. + 16E\left(r s_1 s_2\right) + E\left(6s_1^2\right) + E\left(6s_2^2\right) - 4E\left(3s_1^2\right) - 4E\left(3s_2^2\right) + 6E\left(s_1^2\right) + 6E\left(s_2^2\right) - 7\right]$	
Asymmetric	**Expectation Value**	
$C(2, 1)$	$\left(e^{s_1^2} - 1\right)^{-1}\left(e^{s_2^2} - 1\right)^{-1/2}\left(E\left(s_1^2 + 2r s_1 s_2\right) - 2E\left(r s_1 s_2\right) - E\left(s_1^2\right) + 2\right)$	
$C(3, 1)$	$\left(e^{s_1^2} - 1\right)^{-3/2}\left(e^{s_2^2} - 1\right)^{-1/2}\left(E\left(3s_1^2 + 3r s_1 s_2\right) - 3E\left(s_1^2 + 2r s_1 s_2\right)\right.$ $\left. + 3E\left(r s_1 s_2\right) - E\left(3s_1^2\right) + 3E\left(s_1^2\right) - 3\right)$	
$C(4, 1)$	$\left(e^{s_1^2} - 1\right)^{-2}\left(e^{s_2^2} - 1\right)^{-1/2}\left(E\left(6s_1^2 + 4r s_1 s_2\right) - 4E\left(3s_1^2 + 3r s_1 s_2\right)\right.$ $\left. + 6E\left(s_1^2 + 2r s_1 s_2\right) - 4E\left(r s_1 s_2\right) - E\left(6s_1^2\right) + 4E\left(3s_1^2\right) - 6E\left(s_1^2\right) + 4\right)$	
$C(3, 2)$	$\left(e^{s_1^2} - 1\right)^{-3/2}\left(e^{s_2^2} - 1\right)^{-1}\left(E\left(3s_1^2 + s_2^2 + 6r s_1 s_2\right) - 3E\left(s_1^2 + s_2^2 + 4r s_1 s_2\right)\right.$ $- 2E\left(3s_1^2 + 3r s_1 s_2\right) + 6E\left(s_1^2 + 2r s_1 s_2\right) + 3E\left(s_2^2 + 2r s_1 s_2\right)$ $\left. - 6E\left(r s_1 s_2\right) + E\left(3s_1^2\right) - 3E\left(s_1^2\right) - E\left(s_2^2\right) + 4\right)$	
$C(4, 2)$	$\left(e^{s_1^2} - 1\right)^{-2}\left(e^{s_2^2} - 1\right)^{-1}\left(E\left(6s_1^2 + s_2^2 + 8r s_1 s_2\right) - 4E\left(3s_1^2 + s_2^2 + 6r s_1 s_2\right)\right.$ $+ 6E\left(s_1^2 + s_2^2 + 4r s_1 s_2\right) - 2E\left(6s_1^2 + 4r s_1 s_2\right) + 8E\left(3s_1^2 + 3r s_1 s_2\right)$ $- 12E\left(s_1^2 + 2r s_1 s_2\right) - 4E\left(s_2^2 + 2r s_1 s_2\right) + 8E\left(r s_1 s_2\right) + E\left(6s_1^2\right)$ $\left. + 6E\left(s_1^2\right) - 4E\left(3s_2^2\right) + E\left(s_2^2\right) - 5\right)$	

Table 5. Correlation coefficients $c(p, q)$ of (Scaled H)^p and (Scaled W)^q.

Correlation of Order (p, q)	ANSUR Male (Nm = 4082)		ANSUR Female (Nf = 1986)	
Symmetric	**Theory**	**Empirical**	**Theory**	**Empirical**
$c(1, 1)$	0.4689 ± 0.0176	0.4689	0.5359 ± 0.0259	0.5335
$c(2, 2)$	1.4805 ± 0.0930	1.5016	1.6239 ± 0.1476	1.6273
$c(3, 3)$	5.6321 ± 1.0145	5.3520	6.6972 ± 1.6862	6.2098
$c(4, 4)$	37.4678 ± 17.6228	29.1619	45.9161 ± 30.6070	32.5774
Asymmetric				
$c(2, 1)$	0.0733 ± 0.0386	0.0360	0.0539 ± 0.0594	0.0909
$c(3, 1)$	1.4246 ± 0.0972	1.4258	1.6303 ± 0.1514	1.6475
$c(4, 1)$	0.6653 ± 0.3039	0.2072	0.7936 ± 0.4802	0.4785
$c(5, 1)$	7.3582 ± 1.0496	7.1497	8.4395 ± 1.6738	8.0036
$c(3, 2)$	0.8319 ± 0.2723	0.5211	0.4561 ± 0.4427	0.6455
$c(4, 2)$	6.0898 ± 0.8993	5.8583	6.6411 ± 1.4904	6.6180
$c(4, 3)$	7.3772 ± 3.5970	2.9937	3.7538 ± 6.1366	4.3204
$c(5, 3)$	34.9370 ± 14.1979	28.1909	35.8100 ± 24.7631	32.5825

is a measure of the difference between the joint characteristic function (CF) [28]

$$g_{X,Y}(t,s) = \iint f_{X,Y}(x,y)\exp\big(i(tx+sy)\big)\mathrm{d}x\mathrm{d}y \tag{39}$$

and the product of the marginal CFs

$$\begin{aligned} g_X(t) &= \int f_X(x)\exp(itx)\mathrm{d}x \\ g_Y(s) &= \int f_Y(x)\exp(isy)\mathrm{d}y \end{aligned} \tag{40}$$

As indicated in Equations (39) and (40), the CF g_Z is the Fourier transform of the corresponding probability density f_Z of some specific random variable or set of random variables Z.

The actual evaluation of $V(X,Y)$ in Equation (38), together with its properties and associated theorems, is given in Ref. [12]. For the purposes of this paper, sufice it to say that $V(X,Y)$ involves a weighted integral of $\left|g_{X,Y}(t,s)-g_X(t)g_Y(s)\right|^2$ over the Fourier coordinates t, s. The associated quantity of distance variance (dVar), given by $V(X,X)$, would then be the same weighted integral over $\left|g_{X,X}(t,s)-g_X(t)g_X(s)\right|^2$ with analogous expressions for $V(Y,Y)$. The distance correlation (dCor), expressed by $R(X,Y)$, is then defined in terms of dCov and dVar by the relation

$$R(X,Y) \equiv \frac{\left|V(X,Y)\right|}{\sqrt{\left|V(X,X)\right|\left|V(Y,Y)\right|}}. \tag{41}$$

Equation (41) resembles in form Equation (12) or Equation (14) for the Pear-

son correlation coefficient, but its properties are significantly different, as well as its empirical evaluation. The author is unaware of any theoretical derivations of the probability density function or statistical moments of $R(X,Y)$. However, if X and Y are standard normal variables, then $R(X,Y)$ has been evaluated in the following closed form [12]

$$R^{(N,N)}(X,Y)=\frac{\rho\arcsin(\rho)+\sqrt{1-\rho^2}-\rho\arcsin(\rho/2)-\sqrt{4-\rho^2}+1}{1+(\pi/3)-\sqrt{3}} \tag{42}$$

where the superscript (N, N) signifies the special case of bivariate standard normality, and ρ is the associated Pearson correlation coefficient.

4.1. Statistical Application of Distance Correlation (dCor) to Height and Weight

The procedure for applying dCor to a statistical system is as follows: Given samples (x_i,y_i) for $i=1,\cdots,n$ of two random variables X and Y, construct the statistic

$$A_{k,l}=a_{k,l}-a_{k\bullet}-a_{\bullet l}+a, \tag{43}$$

where $(k,l=1,\cdots,n)$ and

$$\begin{cases} a_{k,l}=|x_k-x_l|, & a_{k\bullet}=\dfrac{1}{n}\displaystyle\sum_{l=1}^{n}a_{k,l} \\ a_{\bullet l}=\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}a_{k,l}, & a=\dfrac{1}{n^2}\displaystyle\sum_{k,l=1}^{n}a_{k,l} \end{cases} \tag{44}$$

and the associated statistic

$$B_{k,l}=b_{k,l}-b_{k\bullet}-b_{\bullet l}+b \quad , \tag{45}$$

where

$$\begin{cases} b_{k,l}=|y_k-y_l|, & b_{k\bullet}=\dfrac{1}{n}\displaystyle\sum_{l=1}^{n}b_{k,l} \\ b_{\bullet l}=\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}b_{k,l}, & b=\dfrac{1}{n^2}\displaystyle\sum_{k,l=1}^{n}b_{k,l} \end{cases} \tag{46}$$

The squares of the empirical dCov and dVar are then given by

$$V_n^2(X,Y)=\frac{1}{n^2}\sum_{k,l=1}^{n}A_{k,l}B_{k,l} \tag{47}$$

$$\begin{aligned} V_n^2(X,X)&=\frac{1}{n^2}\sum_{k,l=1}^{n}A_{k,l}^2 \\ V_n^2(Y,Y)&=\frac{1}{n^2}\sum_{k,l=1}^{n}B_{k,l}^2 \end{aligned} \tag{48}$$

from which follows the square of the empirical dCor

$$R_n^2(X,Y)=\frac{V(X,Y)^2}{\sqrt{V_n^2(X,X)V_n^2(Y,Y)}} \tag{49}$$

corresponding to Equation (41). The statistic $R_n(X,Y)$ 1) ranges between 0

and 1, 2) is 0 only if X and Y are independent, and 3) approaches the theoretical dCor $R(X,Y)$ in the limit of infinite sample size n [12].

4.2. Distance Correlation Test of the ANSUR Data

In the sequence of analyses of the ANSUR data to follow, the pair of variables (X,Y) was taken to be the scaled sets 1) (H,W), 2) $(\ln H,\ln W)$, and 3) $\left(\ln H,(\ln W)_{\text{red}}\right)$. It is to be recalled that a "scaled" variable is in dimensionless form of zero mean and unit variance. The subscript "red" in set (3) indicates that the scaled variates of $\ln W$ were *reduced* by subtraction of the regression of $\ln W$ on $\ln H$. The reason for this reduction and the way it was implemented will be clarified shortly.

The algorithm (Equations (43) to (48)) leading to Equation (49) is straightforward to implement by computer. However, for sample sizes on the order of thousands, the computation time is impractically long. To circumvent this difficulty, a sampling procedure analogous to bootstrapping [34] [35] was employed.

Each bivariate pair (h_i,w_i) of height and weight in the ANSUR data set is labeled by an index i, referred to here as the "participant number", that ranges from 1 to the full sample size n, which is $n_m=4082$ for the male cohort and $n_f=1986$ for the female cohort. Participants in the survey were apparently measured and recorded in random order, as shown in **Figure 11**, which displays scatter plots by gender of the scaled height and weight vs participant number. Although *histograms* of the variates, analyzed in part I [11], are precisely matched by lognormal distributions, the *point density* plots in **Figure 11** are well represented by uniform distributions across the entire range of participants. In other words, each vertical slice of points of sufficient width contains a statistical spread of variates equivalent to any other vertical slice of the same width. Given this uniform density, the ranges n_m and n_f were respectively partitioned into 20 and 10 subgroups of 200 participants each, as shown in **Figure 12**. Distance correlation of height and weight was then evaluated for 50 consecutive participants in each subgroup, starting at the participant numbers marked by diamond plotting symbols in the figure.

For example, $dCor_1$ was calculated from participants [200 - 249], $dCor_2$ from partipants [400 - 449], and so on up to $dCor_{20}$ from participants [4000 to 4049] for males and up to $dCor_9$ from participants [1800 to 1849] for females. As with standard bootstrapping, this method of calculating dCor by repeated sampling not only circumvented what otherwise would have been an excessively long computer calculation, but it also provided a vector of dCor values from which to estimate dCor uncertainty in the absence of a known statistical distribution.

The results of the analyses of distance correlation are summarized in **Table 6**.

Section I of the table records the distance correlation of the scaled variables H and W. Particularly striking is the close agreement between the empirical values

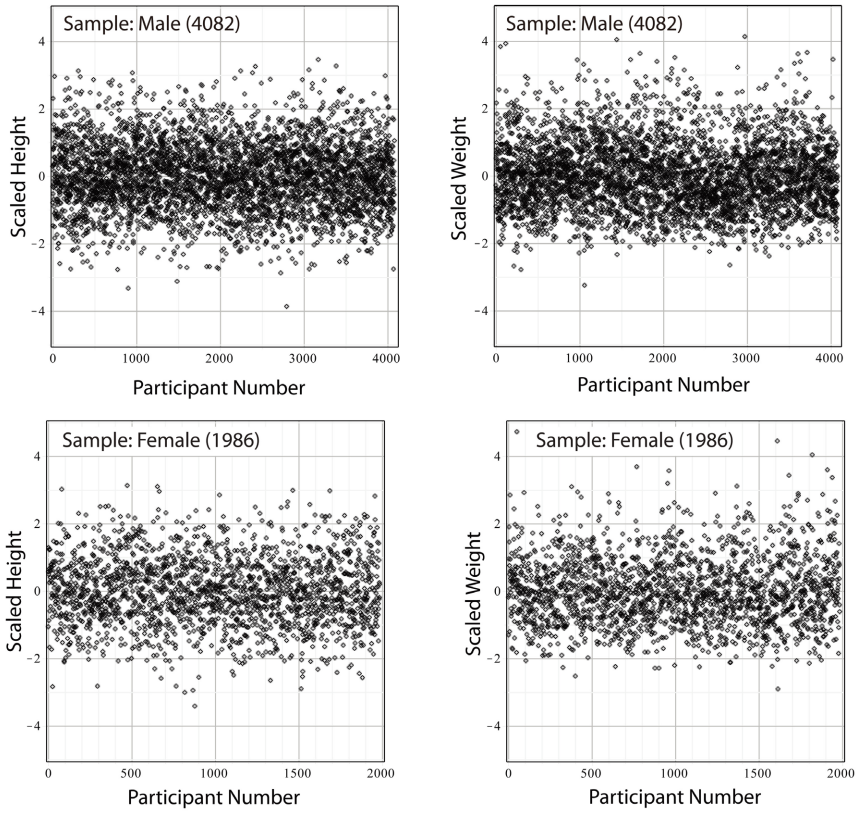
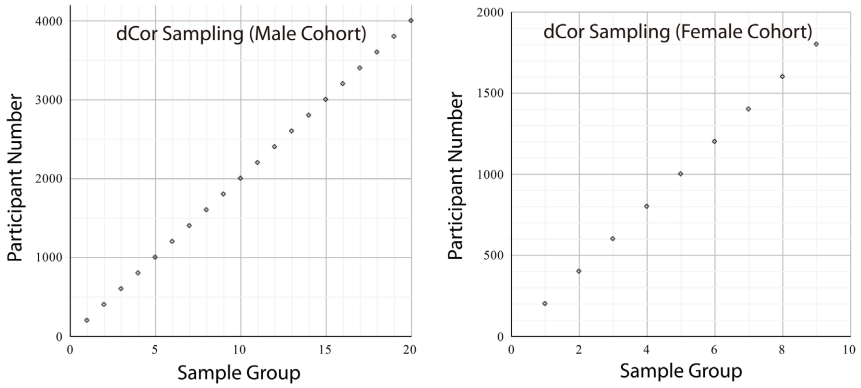


Figure 11. Plots of height (left panels) and weight (right panels) of individual participants in the ANSUR sample: male cohort (top panels); female cohort (bottom panels).



(a) (b)

Figure 12. Illustration of the resampling strategy for calculation of distance correlation of participants' height and weight in the ANSUR population. Diamond plotting symbols mark participant numbers which begin each subgroup of 50 participants to be sampled over the range (a) 200 to 4000 (male cohort), (b) 200 to 1850 (female cohort).

obtained from the ANSUR sample (columns 2 and 3 for male and female cohorts, respectively) and values created by computer simulation using a bivariate lognormal RNG (columns 4 and 5 for the closely corresponding sample sizes 4000 and 2000, respectively). Numerical values designated by "rho" at the top of columns 4 and 5 are the correlation parameters supplied to the RNG. These values

Table 6. Test of nonlinear correlations by means of distance correlation: $\mathrm{dCor}(X,Y)$.

Variables (X, Y)	ANSUR Male (4082)	ANSUR Female (1986)	Simulation (4000) rho = 0.4716	Simulation (2000) rho = 0.5387
(I) $\frac{H-\mu_H}{\sigma_H}, \frac{W-\mu_W}{\sigma_W}$				
Mean dCor	0.4725	0.5004	0.4428	0.5099
SE dCor	0.0259	0.0285	0.0231	0.0243
Pearson ρ	0.4689	0.5335	0.4584	0.5359
Theoretical dCor	0.4248	0.4860	0.4150	0.4883
(II) $\frac{\ln H-m_H}{s_H}, \frac{\ln W-m_W}{s_W}$				
Mean dCor	0.4856	0.5027	0.4443	0.5032
SE dCor	0.0272	0.0292	0.0228	0.1186
Pearson r	0.4716	0.5387	0.4604	0.5030
Theoretical dCor	0.4273	0.4910	0.4169	0.4570
(III) $\frac{\ln H-m_H}{s_H}, \frac{\ln W-m_W}{s_W} - r\left(\frac{\ln H-m_H}{s_H}\right)$			Simulation (4000) rho = 0	Simulation (2000) rho = 0
A. Repeated Sampling of Total Population (Sample Size 50)				
Mean dCor	0.2475	0.2454	0.2275	0.2382
SE dCor	0.0119	0.0104	0.0081	0.0112
Pearson Corr Coeff	-4.1118×10^{-11}	-4.1011×10^{-11}	0.0171	0.0124
Theoretical dCor	0	0	0.0152	0.0110
B. Single Sample of Sub Population (Sample Size 1000)				
Mean dCor	0.0611	0.0676	-	0.0599
Pearson Corr Coeff	−0.0263	−0.0219	-	0.0213
Theoretical dCor	0.0234	0.0195	-	0.0190

(from Table 1) correspond to the empirical Pearson correlation coefficients of the variables $(\ln H, \ln W)$. The empirical Pearson correlation coefficients ρ (columns 2 and 3) of the variables (H,W) are closely matched by the corresponding values (columns 4 and 5) obtained by computer simulation. In short, the dCor values of Section I are consistent with attributing the *entire* correlation of height and weight, including any nonlinear contributions, to the bivariate lognormal distribution.

Section II of the table records distance correlation of the scaled variables $(\ln H, \ln W)$. Computer simulated populations of sizes approximating the male and female ANSUR cohorts were generated by use of bivariate normal RNGs.

Agreement between empirical and corresponding computer simulated values is again very close, especially for the female cohort. The outcome indicates that the full correlation of $\ln H$ and $\ln W$ is attributable to the bivariate normal distribution which, in turn, derives from the parent lognormal distribution function.

The distance correlation procedure tests random vectors for independence irrespective of their specific distributions, provided the first moments are finite [13]. It is a nonparametric test that can reveal correlations even when the Pearson correlation coefficient is null. However, if the Pearson correlation of two *normal* vectors is null, then those vectors are *fully independent—i.e.* there is no latent nonlinear correlation. Section III of the table exploits this point to ascertain whether the correlation between height and weight exhibited in **Figure 1** and **Figure 2** have a nonlinear contribution *not* attributable to the parent lognormal or derived normal distributions.

The basic idea is to subtract from the ordinate of each point in the scatter plots in **Figure 2** the corresponding ordinate of the line of regression. For correlations of scaled standard normal variables (X,Y), the line of regression takes the simple form $y = rx$, where r is the slope of the line and is equal to the Pearson correlation coefficient. A scatter plot is then made of the *reduced* scaled log weight

$$(\ln W)_{red} \equiv (\ln W - m_W)/s_W - r(\ln H - m_H)/s_H \tag{50}$$

against the scaled log height $(\ln H - m_H)/s_H$, as shown in **Figure 13**. The scatter plots (black points) for male (left) and female (right) ANSUR cohorts exhibit the same isotropic patterns (apart from fluctuations) as the simulated scatter plot for null correlation (rho = 0) in **Figure 5**. Quantitatively, the slopes of the lines of regression (dashed red) of the reduced plots are respectively -4.1118×10^{-11} (male cohort) and -4.1011×10^{-11} (female cohort). In other words, removal of the lines of regression from the empirical scatter plots of

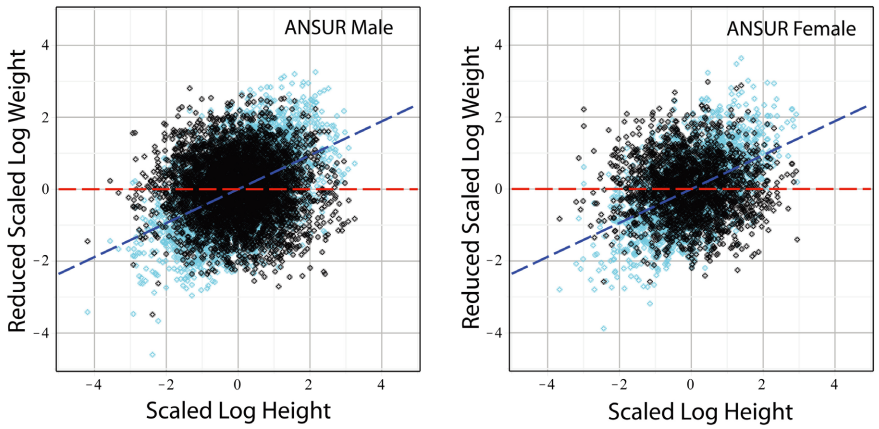


Figure 13. Scatter pattern (black points) and associated line of regression (dashed red) of zero slope signifying a null Pearson correlation of log weight and log height when the log weight variates were reduced by corresponding values of the line of regression (dashed blue) of the original scatter plot (cyan points).

scaled $(\ln H, \ln W)$ has resulted in two normal random vectors of null Pearson correlation coefficient—and therefore presumably statistically independent. For comparison, the reduced scatterplots (black points) in Figure 13 are superposed on the original scatterplots (cyan points) of Figure 2 with their lines of regression (dashed blue).

The dCor values in Section III provide quantitative confirmation of the statistical independence of height and weight upon removal of the Pearson linear correlation. In the subsection A based on repeated sampling of the entire population in samples of size 50, the empirical dCor values (approximately 0.25) agree closely with dCor values (approximately 0.23 to 0.24) produced by two independent standard normal RNGs. The small deviations are attributable in part to the fact that the resulting Pearson correlation coefficients of the simulated populations were not precisely 0, but in the range 0.01 to 0.02, a consequence of the fluctuations intrinsic to finite sampling.

Potentially problematic is the discrepancy between the empirical (as well as simulated) dCor values obtained by repeated sampling and the values predicted by Equation (42), which should be close to 0 for two independent normal random variables. However, Equation (42) is strictly valid only in the limit of an infinite population. Subsection B, based on single sampling of a much larger subpopulation of 1000 participants, shows that empirical dCor values dropped to approximately 0.06, in much closer agreement with Equation (42). Evaluation of the distance correlation of a sample of 1000 required computation times longer than 8 hours. Thus, to test rigorously whether dCor approaches 0 asymptotically as a function of sample size would require impractically long computation times.

Altogether, the three sections of Table 6 consistently support the conclusion that the observed correlation between height and weight can be accounted for entirely by a bivariate lognormal distribution. In other words, the five parameters defining the bivariate lognormal distribution of height and weight suffice to predict any measureable function or test of the correlation of height and weight of a healthy adult human population.

5. Marginal Statistics of ANSUR Height and Weight Data

Previous sections concentrated on the bivariate lognormal correlation of height and weight. This section examines the marginal statistics of H and W, which are predicted to follow the respective univariate lognormal distributions $\Lambda\left(m_H, s_H^2\right)$ and $\Lambda\left(m_W, s_W^2\right)$, and on $\ln H$ and $\ln W$, which are predicted to follow the respective univariate normal distributions $N\left(m_H, s_H^2\right)$ and $N\left(m_W, s_W^2\right)$ as discussed in Section 1.1.

Table 7 summarizes the outcomes of chi-square tests of fitness of the four variables H, W, $\ln H$, $\ln W$ to their respective distributions, identified explicitly in column 3. As a reminder, the chi-square statistic χ_ν^2, in column 7 of the table, is determined empirically from the relation

$$\left(\chi_\nu^2\right)_{emp} = \frac{1}{\kappa}\sum_{i=1}^{\kappa}\frac{\left(O_i - E_i\right)^2}{E_i} \tag{51}$$

Table 7. Chi-square tests of goodness of fit of ANSUR height and weight.

Test Variable	Cohort: M (4082) F (1986)	Distribution	Critical Value	d.o.f	*P*-Value	χ_ν^2 Statistic
Height H	Male	$\Lambda\left(0.5624, 0.0390^2\right)$	82.529	63	0.5952	59.686
	Female	$\Lambda\left(0.4869, 0.0394^2\right)$	60.481	44	0.1754	52.600
Weight W	Male	$\Lambda\left(4.4351, 0.1654^2\right)$	82.529	63	0.9694	43.720
	Female	$\Lambda\left(4.2030, 0.1604^2\right)$	60.481	44	0.6225	40.498
Log Height $\ln H$	Male	$N\left(0.5624, 0.0390^2\right)$	82.529	63	0.3626	66.339
	Female	$N\left(0.4869, 0.0394^2\right)$	60.481	44	0.2700	49.287
Log Weight $\ln W$	Male	$N\left(4.4351, 0.1654^2\right)$	82.529	63	0.7587	54.828
	Female	$N\left(4.2030, 0.1604^2\right)$	60.481	44	0.3201	47.826

where κ is the number of test categories (bins), O_i is the observed value in the i^{th} bin, and E_i is the expected value in the i^{th} bin. The subscript ν (Greek nu) is the number of degrees of freedom (d.o.f.) in column 5 equal to $\kappa - 1$. The chi-square tests were implemented with the *Maple* Statistics Package, which determined the number of bins as the integer closest to the square root of the sample size. The critical values in column 4 of the table are the values of χ_ν^2 resulting in *P*-values of 5%, which is the conventional threshhold of statistical significance; *i.e.* a tested hypothesis is deemed unsupported if the *P*-value is below threshhold. A *P*-value is the probability of obtaining a test result at least as extreme as the observed result χ_{obs}^2, and is calculated from the expression

$$P = \int_{\chi_{\text{obs}}^2}^{\infty} p_{\chi_\nu^2}(z)\,\mathrm{d}z \tag{52}$$

with chi-square PDF

$$p_{\chi_\nu^2}(z) = \frac{z^{\frac{1}{2}\nu - 1}\mathrm{e}^{-z/2}}{2^{\nu/2}\Gamma(\nu/2)}. \tag{53}$$

The outcomes summarized in the table show that all 8 propositions (the distributions of 4 variables of 2 genders) passed their respective chi-square tests with *P*-values far above threshhold. This means that the propositions cannot be rejected on the basis of these tests. It does not necessarily mean, however, that the propositions are true.

For further confirmation, consider again the information in Table 1. In the first section of the table, empirical values of the mean, standard deviation, skewness, and kurtosis of ANSUR heights H and weights W are compared with corresponding values predicted by lognormal expressions (7) to (10), based on the parameters in the second section of the table, derived from the variates of $\ln H$

and $\ln W$. Agreement of experiment and theory is seen to be within 1 standard error (se) in most cases. Table 1 employed the following published estimators of the standard errors for skewness and kurtosis [36]

$$se(n)_{Sk} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \tag{54}$$

$$se(n)_{K} = 2\sqrt{\frac{6n(n-1)^2}{(n-2)(n+5)(n^2-9)}}. \tag{55}$$

In the second section of Table 1 the empirical skewness of both $\ln H$ and $\ln W$ for both cohorts are close to zero, as expected for normal variables. Likewise, the empirical kurtosis is very close to 3, as expected for normal variables. Skewness is a measure of the asymmetry of a distribution about the mean. Kurtosis (from the Greek root for "bulging") is a measure of the curving or arching of the *tails* of a distribution; in other words, kurtosis is an indicator of the extent of outliers, relative to the normal distribution.

Ordinarily, standardized statistical moments employed in physical science and medicine include at most only the first four orders (mean, variance, skewness, kurtosis). Beyond these, higher standardized moments, such as "hyperskewness" and "hyperkurtosis" [37], are rarely used in the author's experience, presumably because they are less readily interpetable as well as have greater measures of uncertainty for a given sample size. Nevertheless, in testing a proposed statistical distribution, it is useful to examine these higher moments, particularly if the validity of all lower moments has been confirmed.

In the terminology and notation of this paper, the hyperstatistic $S_p(X)$ of the random variable X is the p^{th} standardized central moment defined by the relation

$$S_p(X) \equiv \frac{\left\langle (X-\mu_X)^p \right\rangle}{\left\langle (X-\mu_X)^2 \right\rangle^{p/2}} \equiv \frac{\mu_p(X)}{\mu_2^{p/2}(X)} \tag{56}$$

where μ_X is the mean of X, and $\mu_p(X)$ is the mean of the random variable

$$m_p(X) \equiv (X-\mu_X)^p \tag{57}$$

referred to as the p^{th} central moment. To simplify symbolic notation in the ensuing text, the argument (X) will be omitted whenever the context is clear.

Substitution in relation (56) of the univariate lognormal PDF, mean, and variance leads to the operational expression

$$S_p(X) = \frac{\int_{-\infty}^{\infty} \left(e^{su} - e^{\frac{1}{2}s^2} \right)^p p_X^{\Lambda}(u)\,du}{\left(e^{2s^2} - e^{s^2} \right)^{\frac{p}{2}}} \tag{58}$$

where s is the standard deviation of the variable $\ln X$. Skewness and kurtosis correspond respectively to p = 3, 4. Table 8 lists the theoretical expressions for hyperstatistics of order 1 through 6 derived from relation (58).

Table 8. Theoretical hyperstatistics of univariate lognormal distribution. (Notation: $E(x) \equiv \exp(x)$).

Statistic	Expectation Value
$S_1(X)$	0
$S_2(X)$	1
$S_3(X)$	$\left(\mathrm{e}^{2s^2}-\mathrm{e}^{s^2}\right)^{\frac{3}{2}}\left(E\left(\frac{9}{2}s^2\right)-3E\left(\frac{5}{2}s^2\right)+2E\left(\frac{3}{2}s^2\right)\right)$
$S_4(X)$	$E\left(4s^2\right)+2E\left(3s^2\right)+3E\left(2s^2\right)-3$
$S_5(X)$	$\left(\mathrm{e}^{2s^2}-\mathrm{e}^{s^2}\right)^{\frac{5}{2}}\left(E\left(\frac{25}{2}s^2\right)-5E\left(\frac{17}{2}s^2\right)+10E\left(\frac{11}{2}s^2\right)-10E\left(\frac{7}{2}s^2\right)+4E\left(\frac{5}{2}s^2\right)\right)$
$S_6(X)$	$E\left(12s^2\right)+3E\left(11s^2\right)+6E\left(10s^2\right)+10E\left(9s^2\right)+15E\left(7s^2\right)$ $+10E\left(6s^2\right)-15E\left(4s^2\right)-20E\left(3s^2\right)-15E\left(2s^2\right)+5$

Table 9 summarizes the empirical results for hyperstatistics of orders $p = 5, 6$ for the variables H and W of the ANSUR population. The empirical entries (column 3) are the sample statistics

$$S_p(X)_{emp} = \frac{\overline{m_p}}{\left(\overline{m_2}\right)^{\frac{p}{2}}} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^p \Bigg/ \left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\right)^{\frac{p}{2}} \tag{59}$$

where $\overline{x}$ is the sample mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{60}$$

and the numerator

$$\overline{m_p} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^p \tag{61}$$

is the expectation of the sample p^{th} central moment (57). Theoretical entries in column 4 were calculated from Equation (58). Overall, agreement between theory and experiment appears reasonably close, but several statistics show what may be significant deviations. To ascertain whether any deviation between theory and experiment is statistically significant requires knowing the standard error of the mean statistic, but the author is unaware of any published expressions for the distributions or standard errors of hyperskewness and hyperkurtosis. To estimate the pertinent standard errors, three independent approaches were taken.

The first approach was to use the approximations of error propagation theory [38] together with expressions for variance and covariance of central moments in Chapter 10 of Ref. [28] to derive an estimate of the variance of hyperstatistic $S_p(X)$

Table 9. Test of hyperstatistics of height and weight of ANSUR population.

Hyperstatistic $S_p(X)$ $p = (5, 6)$ $X = (H, W)$	Cohort M: 4082 F: 1986	Empirical	Theory	Simulation M: 4082 F: 1986
$S_5(H) = \left\langle (H-\mu_H)^5 \right\rangle / \sigma_H^5$				
Mean $S_5(H)$	male	0.9679	1.1784	0.5997 1.0966 1.1013
	female	0.4836	1.1907	0.6304 1.2628 0.5685
Standard Error $S_5(H)$	male	0.3983	0.4518	0.5016 | 0.1664
	female	0.4835	0.6487	0.6943 | 0.2218
$S_5(W) = \left\langle (W-\mu_W)^5 \right\rangle / \sigma_W^5$				
Mean $S_5(W)$	male	4.8791	5.6489	7.6398 4.7215 5.6557
	female	6.5463	5.4361	4.1332 3.7024 5.9618
Standard Error $S_5(W)$	male	0.5518	1.0841	2.9183 | 0.8604
	female	1.3757	1.4869	2.2594 | 0.6926
$S_6(H) = \left\langle (H-\mu_H)^6 \right\rangle / \sigma_H^6$				
Mean $S_H^{(6)}$	male	15.6714	15.5060	15.9807 13.5489 14.4296
	female	14.3941	15.5165	13.1332 15.0664 13.1147
Standard Error $S_H^{(6)}$	male	1.0394	1.4580	2.8475 | 0.7108
	female	1.1307	2.0974	1.9517 | 0.6475
$S_6(W) = \left\langle (W-\mu_W)^6 \right\rangle / \sigma_W^6$				
Mean $S_6(W)$	male	21.3981	25.4523	38.8263 21.0697 26.7901
	female	29.0534	24.7436	18.5144 17.0547 24.5282
Standard Error $S_6(W)$	male	1.9545	6.0795	17.7566 | 5.2329
	female	6.0559	8.1822	7.4735 | 2.2870

$$\operatorname{var}\left(S_p\right)=\frac{\operatorname{var}\left(m_p\right)}{\mu_2^k}+\frac{p^2\mu_p^2}{4\mu_2^{k+2}}\operatorname{var}\left(m_2\right)-\frac{p\mu_k}{\mu_2^{k+1}}\operatorname{cov}\left(m_k,m_2\right) \tag{62}$$

in which

$$\operatorname{var}\left(m_p\right)=\mu_{2p}-\mu_p^2+p^2\mu_2\mu_{p-1}^2-2p\mu_{p-1}\mu_{p+1} \tag{63}$$

and

$$\operatorname{cov}\left(m_p,m_q\right)=\mu_{p+q}-\mu_p\mu_q+pq\mu_2\mu_{p-1}\mu_{q-1}-p\mu_{p-1}\mu_{q+1}-q\mu_{p+1}\mu_{q-1}. \tag{64}$$

The standard error is then

$$se\left(S_p\right)=\sqrt{\frac{\operatorname{var}\left(S_p\right)}{n}}. \tag{65}$$

Theoretical evaluations of Equation (65) by means of the univariate lognormal PDF are recorded in column 4.

The second approach was to evaluate the means of all moments in Equation (62) by their sample expectations given by Equation (61). The resulting empirical standard errors are recorded in column 3.

In the third approach three independent simulations of the male and female ANSUR populations were made with correlated lognormal RNGs representing the variables H and W, from which empirical values of the individual hyperstatistics were obtained. The three values for each hyperstatistic (p = 5, 6) of the two variables (H, W) for the two genders (M, F) are listed in column 5 in the cells for the mean of $S_p(X)$. The sample standard error of each set of 3 independent means was calculated from the relation

$$se\left(S_p(X)\right)=\frac{1}{\sqrt{n(n-1)}}\sqrt{\sum_{k=1}^{3}\left(S_p(X)_k-\overline{S_p(X)}\right)^2} \tag{66}$$

and recorded as the first entry in the cells for standard error in column 5. Note that n is Equation (66) is 3 and not the ANSUR population size of 4082 males or 1986 females. Also, because a sample size of 3 is statistically small, one uses the *unbiased* sample estimate of variance in Equation (66) in which the denominator is $n-1$ [39]. The second entry (separated from the first by a vertical bar) is the *difference* between the largest and smallest of the three estimated means of each $S_p(X)$.

In examining the three sets of estimated standard errors, one sees that they are approximately of the same magnitude, and that the empirical values of the hyperstatistics agree with theory within $\pm 2se$ for at least one of the three estimates. However, one glaring exception is the value of $se\left(S_6(W)_{\mathrm{M}}\right)$ for the male cohort obtained by simulation, which is much larger than the other standard errors for the same statistic. This occurs because of what appears to be an exceptionally high value (~38) of the mean of $S_6(W)_{\mathrm{M}}$ returned by one of the simulations. This occurrence raises an important conceptual issue that calls for caution when estimating standard errors under conditions where the exact statistical distribution is unknown.

As pointed out in Chapter 10 of Ref [28], the approximations based on error propagation theory (or some variant thereof) give a valid measure of precision provided that the distribution of the statistic approaches normality in the limit of a sufficiently large sample. This is not the case for higher orders of the hyperstatistics. The mere fact that a mean value of the statistic $S_6(W)_{\mathrm{M}}$ can arise *within just* 3 *simulations* that is 6 times the theoretical and 17 times the empirical standard errors of error propagation theory shows that such outlying values occur with much higher probabilities than would be predicted by a normal distribution. Under such circumstances, the appropriate way to proceed is to determine whether the empirical mean values of the hyperstatistics fall within the range between the lowest and highest corresponding statistics obtained by simulation; in other words to rely on a kind of Monte Carlo validation. Evaluated this way, all the empirical hyperstatistics in **Table 9** are seen to be consistent with theory.

Taken together, the chi-square tests and agreement of theoretical and empirical moments (or functions of moments) up to the 6th power of the variables support the propositions that H and W are marginally lognormal variables.

6. Implications for Body Mass Index (BMI)

The BMI, defined by the random variable B in Equation (1), has long been used as a measure of obesity and a risk factor for associated diseases under the assumption that it is correlated with weight but largely independent of height. In Section 2 the conditional expectation function of weight, given height, was derived on the basis of lognormal theory and shown to be very nearly a quadratic power law $W(H) \propto H^2$ (27) for the ANSUR male and female cohorts. If weight varies as the square of height, then the BMI (1) would be unaffected by variations in height, or, in other words, statistically independent of height.

Figure 14 provides additional justification of the BMI assumption. The left panels of the figure display scatter plots of the correlation of scaled BMI and scaled height for the ANSUR male cohort (top), ANSUR female cohort (middle), and RNG simulated female cohort (bottom). The isotropic patterns (apart from fluctuations) are very close to what are expected for the correlation of independent vectors, as shown in the first panel of **Figure 5**. Quantitatively, the lines of regression (dashed red for ANSUR, dashed blue for simulation) yield Pearson correlation coefficients 2.096×10^{-3} (male), 5.406×10^{-2} (female), and 6.700×10^{-2} (simulation). The three sets of variables are not standard normal variables, so a null Pearson correlation does not necessarily imply total independence.

However, evaluation of the distance correlation of the scaled variables (H,B) for the male and female cohorts (black points) respectively yielded by the method of repeated sampling empirical dCor values of 0.2475 ± 0.0116 and 0.2301 ± 0.0147, which are statistically equivalent to the dCor value for correlation of independent standard normal variables generated by computer simulation

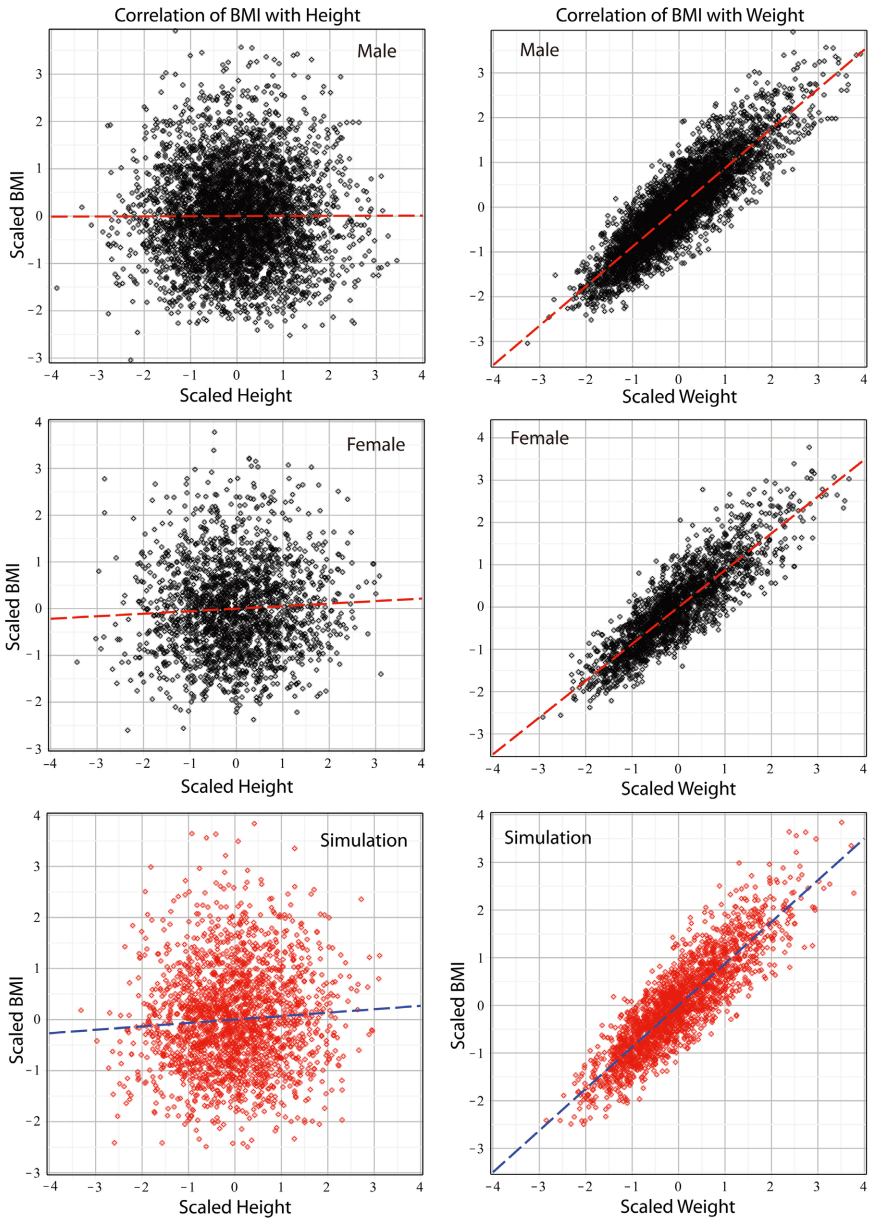


Figure 14. Scatter plots of body mass index (BMI) with height (left panels) and weight (right panels). Patterns in black were calculated from the ANSUR population of male (top panels) and female (middle panels) cohorts. Patterns in red (bottom panels) were simulated for a population of 2000 (corresponding to the size of the female cohort) by means of correlated bivariate lognormal RNGs. The slopes of the lines of regression (dashed red or dashed blue) of the left panels are close to zero, signifying independence of BMI and height. The slope of the lines of regression of the right panels are close to 0.88 for the male cohort and 0.87 for the female cohort.

(red points). It is therefore reasonable to conclude that, if the ANSUR populations can serve as baselines, then the BMI and height of healthy adult male and female populations are largely statistically independent.

By contrast, the right panels of **Figure 14** (black for empirical and red for simulation) show that BMI and weight are strongly linearly correlated, which was

a desirable characteristic of the BMI. The Pearson correlation coefficients are respectively 0.8814 (male cohort), 0.8706 (female cohort), and 0.8755 (simulation of female cohort) as deduced algebraically from Equation (15) or from the slopes of the lines of regression.

7. Computer Simulation of Correlated Lognormal Random Vectors

Throughout this paper computer simulation of correlated random variables (RVs) has been employed for both analytical and graphical comparisons with corresponding empirical results. A brief description of the implementation of these simulations is given in this section.

The essential objective is the simulation of a pair of correlated lognormal RVs of specified parameters (m_1, m_2, s_1, s_2, r). The starting point for the construction is the well-known algebraic identity for decomposition of a general normal variable [16]

$$N_1\left(m_1, s_1^2\right) = m_1 + s_1 N_1(0,1) \tag{67}$$

$$N_2\left(m_2, s_2^2\right) = m_2 + s_2 N_2(0,1) \tag{68}$$

where $N_1(0,1)$ and $N_2(0,1)$ are independent standard normal variables (ISNVs) of mean 0 and variance 1 that serve as basis states. Each ISNV represents a random number generator (RNG) of one's mathematical software. Populations of size n are simulated by creating sets of n variates from each ISNV. The covariance $\left\langle \left(N_1\left(m_1, s_1^2\right) - m_1\right)\left(N_2\left(m_2, s_2^2\right) - m_2\right)\right\rangle$ is theoretically 0 (because $\left\langle N_1(0,1) N_2(0,1)\right\rangle$ is zero) and empirically should approach 0 numerically in the limit of increasing sample size n.

To create a normal RV $N_{2c}\left(m_2, s_2^2, r\right)$ correlated with the RV in Equation (67), one makes the following linear superposition

$$N_{2c}\left(m_2, s_2^2, r\right) = m_2 + r s_2 N_1(0,1) + \sqrt{1-r^2}\, s_2 N_2(0,1). \tag{69}$$

Note that, according to the algebraic rules [16] that govern manipulation of independent normal RVs,

$$aN_1(0,1) + bN_2(0,1) = N_1\left(0, a^2\right) + N_2\left(0, b^2\right) = N\left(0, a^2 + b^2\right), \tag{70}$$

where a and b are constants, one could combine the second and third terms in the right side of Equation (69) to recover the *marginal* distribution represented by Equation (68), since the correlation parameter r drops out. The set of RVs (67) and (69) then comprise a pair of correlated bivariate normal RVs, which, when implemented numerically on a computer, generate the respective variates $\left(y_{1,i}, y_{2,i}\right)$ for $i = 1, \cdots, n$. In the context of simulating samples of human height and weight, these variates are

$$\begin{aligned} y_{1,i} &= \ln\left(h_i\right) \\ y_{2,i} &= \ln\left(w_i\right) \end{aligned} \tag{71}$$

Once one has created the sets of normal variates (71), it remains only to ex-

ponentiate them

$$\begin{aligned} x_{1,i} &= \exp\left(y_{1,i}\right) = h_i \\ x_{2,i} &= \exp\left(y_{2,i}\right) = w_i \end{aligned} \tag{72}$$

to simulate a sample of correlated bivariate lognormal RVs. With some mathematical applications such as *Maple*, one can work with the abstract vectors and simply enter a statement like $X = \exp(Y)$, whereupon the application will know to exponentiate the individual variates as in relation (72). Thus, the bivariate lognormal vectors corresponding to Equations (67) and (69) generated by *Maple* were defined and implemented by the forms

$$\Lambda_1\left(m_1, s_1^2\right) = \exp\left(m_1 + s_1 N_1(0,1)\right) \tag{73}$$

$$\Lambda_{2c}\left(m_2, s_2^2, r\right) = \exp\left(m_2 + r s_2 N_1(0,1) + \sqrt{1-r^2}\, s_2 N_2(0,1)\right) \tag{74}$$

For completeness, a final comment as to the actual nature of the RNGs employed in this paper is called for. All RNGs that employ a mathematical algorithm, in contrast to RNGs based on some random quantum process such as radioactive decay [40], generate pseudo-random numbers. These are sets of numbers that are generated reproducibly from a known starting point (seed value), yet nevertheless pass diverse statistical tests for randomness. The more stringent a test, and the more tests a RNG passes, the better is the RNG.

The MersenneTwister algorithm supplied by the *Maple* RandomTools Subpackage has passed the diehard tests of randomness [41] by G. Marsaglia as well as other tests, and provides numbers that can be considered cryptographically secure [42] [43]. It is safe to accept, therefore, that the independent normal basis states with which the simulation algorithm began and from which the correlated bivariate lognormal distributions were created were for all practical purposes sufficiently uncorrelated.

8. Conclusions and Interpretation

Knowledge of the exact distribution function of a random quantity provides the most complete statistical information attainable about that quantity. This is especially important in regard to anthropometric attributes the statistics of which are essential to clinical medicine and epidemiology.

The fundamental conclusion of this paper is that human height H and weight W in a population of healthy adults are statistically distributed as correlated bivariate lognormal random variables. Moreover, for all practical purposes, this distribution is thought not to be approximate, but empirically rigorous in samples of sufficient size. This means that five measurable parameters, comprising two means $\left(m_H, m_W\right)$, two variances $\left(s_H^2, s_W^2\right)$, and the Pearson linear correlation coefficient (r), suffice to determine *all* statistical attributes (probabilities, moments, correlations) regarding the relation of height and weight in a specified population.

In support of this conclusion, detailed statistical analyses of an extensive anthropometric data base of diverse individuals have shown the following:

- The variates of H and W of both gender cohorts satisfy chi-square tests of fitness to univariate lognormal distributions.
- The variates of $\ln H$ and $\ln W$ of both gender cohorts satisfy chi-square tests of fitness to univariate normal distributions.
- The sample moments of (H,W) up to 6th order are consistent with the lognormal distribution.
- The sample moments of $(\ln H,\ln W)$ up to 6th order are consistent with the normal distribution.
- Theoretical correlation functions $R_{p,q}(r)$ of the normal distribution predict correctly the sample correlation coefficients $r_{p,q}$ of the variates of $(\ln H,\ln W)$.
- Theoretical correlation functions $C_{p,q}(r,s_H,s_W)$ of the lognormal distribution predict correctly the sample correlation coefficients $c_{p,q}$ of the variates of (H,W).
- Computer simulations using correlated lognormal random number generators (RNGs) produce correlation functions and probability density plots of the variables (H,W) that statistically match the corresponding empirical functions and plots.
- Computer simulations using correlated normal (RNGs) produce correlation functions and probability density plots of the variables $(\ln H,\ln W)$ that statistically match the corresponding empirical functions and plots.
- Empirically measured distance correlation (dCor) values of the variables (H,W) agree with dCor values obtained from comparably sized populations simulated by correlated lognormal RNGs.
- Empirically measured distance correlation (dCor) values of the variables $(\ln H,\ln W)$ agree with dCor values obtained from comparably sized populations simulated by correlated normal RNGs.
- Removal of the line of regression in the empirical density plot of $\ln W$ vs $\ln H$ produces an isotropic density with null Pearson correlation coefficient. The density plot, null Pearson correlation coefficient, and dCor values statistically match the corresponding outcomes from two independent normal RNGs.

In short, taken altogether, the preceding extensive set of tests supports the proposition that the distribution and correlation of height and weight of a healthy adult human population are fully accounted for by a bivariate lognormal distribution.

A secondary point worth noting, given the importance of body mass index (BMI) to current medicine and epidemiology, is that the conditional expectation of weight, given height, theoretically derived from the lognormal distribution function yielded functional relations (23), (24) between weight and height. These functions, when evaluated with the lognormal parameters of the ANSUR male and female cohorts, led in both cases to a nearly exact quadratic power law (27), thereby justifying theoretically a long-held assumption underlying the use of

BMI as a risk factor for obesity-related diseases.

In concluding this paper, it is useful to clarify what is meant by an "exact" statistical distribution. In the opinion of the author, who is an atomic and nuclear physicist, statistical distributions in science can arise, broadly speaking, in two ways.

The most fundamental way is as a consequence of a particular dynamical model. In physics, for example, the decay of radioactive nuclei is rigorously accounted for by a binomial probability function, based on a physical model of the independent decay of discrete, uncorrelated nuclei [44]. If the assumption of independence were found to be invalid—and there have been a considerable number of such challenges, only to have been debunked by more careful experiment and analysis [45] [46] [47] [48]—the discovery would have led to deep new insights into the structure and behavior of matter.

The second, less fundamental way, but nevertheless one of practical utility, is by empirical recognition and subsequent verification. To return to the previous physics example, suppose that the phenomenon of radioactive decay was discovered *before* there was any understanding or general acceptance of atoms as discrete units of matter[3]. Then radioactive decay would have been empirically observed to be a Poisson process, and, indeed, the Poisson distribution is widely depicted in books as a rigorous physical law. (See, for example [49].) However, in retrospect, a Poisson process can be interpreted as a degenerate case of a binomial process in the limit of a large number N of radioactive atoms with low probability p of decay, such that Np is the mean number of decays within a specified time interval.

The point of the foregoing examples is this: The rigorously exact distribution (binomial) revealed critical information about the constituents (discrete, independent) of the system. The apparently exact distribution (Poisson) was empirical and utilitarian, but revealed little about the system other than that the decay products were discrete. Under appropriately conceived radiation experiments, the difference between the binomial and Poisson distributions can be observed [50], and the fundamentality of the binomial distribution is established.

In regard to the statistical attributes of human height and weight, the consistency with a correlated bivariate lognormal distribution is, as shown in this paper, so extensive and close, that one must wonder whether it is a rigorously exact consequence of some biophysical mechanism or a limiting case of some other statistical process. How, for example, might a lognormal distribution arise from other distributions?

One such process might entail a random variable X comprising a *product* of some set of arbitrarily distributed random variables, in which case application of the Central Limit Theorem to $\ln X$ could result in a normal distribution. Then the parent variable X would itself be lognormal. It is difficult to conceive in de-

[3]This supposition is actually historically correct. Radioactivity was discovered by Henri Bequerel in 1896, whereas opposition to the existence of atoms by some leading scientists of the day lasted until about 1910.

tail, however, of mechanisms by which real biological processes responsible for human height and weight could engender such a hypothetical X as to produce a correlated bivariate lognormal distribution.

More generally, a lognormal distribution can also arise under circumstances where an intrinsically positive variable has a low mean and high variance, leading to a pronounced skewness. However, any of a large number of other skewed distributions could also arise, so the mechanism is not unique. Moreover, as demonstrated in this paper, whatever mechanism is invoked must produce not only the correct skewness, but also kurtosis and other hyperstatistics as well.

At this stage and until testable mechanisms are proposed, refutation of the exactness of the correlated bivariate lognormal distribution of human height and weight can only come from further detailed statistical analysis of larger populations. And if such future tests further confirm the exactness of the bivariate lognormal relation of height and weight, then, like the example of radioactivity cited above, this knowledge will have revealed something fundamental about the physical processes underlying human development.

Acknowledgements

The author thanks Trinity College for partial support through the research fund associated with the George A. Jarvis Chair of Physics.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Stigler, S.M. (1986) The History of Statistics: The Measurement of Uncertainty Before 1900. Harvard University Press, Cambridge, 265-361.

[2] Bernstein, P.L. (1998) Against the Gods: The Remarkable Story of Risk. Wiley, New York, 152-171. https://doi.org/10.2307/2685740

[3] Porter, T.M. (2004) Karl Pearson: The Scientific Life in a Statistical Age. Princeton University Press, Princeton, 235-239, 249-266. https://doi.org/10.5944/empiria.8.2004.989

[4] Quetelet, L.A.J. (1835) A Treatisse on Man and the Development of His Faculties. Cambridge University Press, Cambridge. https://www.cambridge.org/core/books/treatise-on-man-and-the-development-of-his-faculties/AB13A647A6C8727C06AE5399D7422887

[5] Sager, G. (1987) Relation between Body Height and Weight in Adult Humans. *Gegenbaurs morphologisches Jahrbuch*, **133**, 563-571.

[6] Rahmandad, H. (2014) Human Growth and Body Weight Dynamics: An Integrative Systems Model. *PLOS ONE*, **9**, e114609. https://doi.org/10.1371/journal.pone.0114609 https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0114609

[7] Lettre, G. (2011) Recent Progress in the Study of the Genetics of Height. *Human Genetics*, **129**, 465-472. https://doi.org/10.1007/s00439-011-0969-x

[8] Wikipedia (2022) Body Mass Index. https://en.wikipedia.org/wiki/Body_mass_index

[9] World Health Organization (2021) Obesity and Overweight. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[10] Moody, J.N., *et al.* (2021) Body Mass Index and Polygenic Risk for Alzheimer's Disease Predict Conversion to Alzheimer's Disease. *The Journals of Gerontology Series A Biological Sciences and Medical Sciences*, **76**, 1415-1422. https://doi.org/10.1093/gerona/glab117

[11] Silverman, M.P. and Lipscombe, T.C. (2022) Exact Statistical Distribution of the Body Mass Index (BMI): Analysis and Experimental Confirmation. *Open Journal of Statistics*, **12**, 324-356. https://doi.org/10.4236/ojs.2022.123022

[12] Szekely, G.J., Rizzo, M.L. and Bakirov, N.K. (2007) Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, **35**, 2769-2794. https://doi.org/10.1214/009053607000000505

[13] Szekely, G.J. and Rizzo, M.L. (2009) Brownian Distance Covariance. *The Annals of Applied Statistics*, **3**, 1236-1265. https://doi.org/10.1214/09-AOAS312

[14] Callahan, A. (2021) Is BMI a Scam? *The New York Times*. https://www.nytimes.com/2021/05/18/style/is-bmi-a-scam.html

[15] Gordon, C.C., *et al.* (2014) 2012 Anthropometric Survey of U.S. Army Personnel: Methods and Summary Statistics. Technical Report Natick/TR-15/007, U.S. Army Natick Soldier Research and Engineering Center, Natick. https://www.openlab.psu.edu/ansur2

[16] Silverman, M.P. (2014) A Certain Uncertainty: Nature's Random Ways. Cambridge University Press, Cambridge, 511-514. https://doi.org/10.1017/CBO9781139507370

[17] Forbes, C., Evans, M., Hastings, N. and Peacock, B. (2011) Statistical Distributions. 4th Edition, Wiley, New York, 131-134. https://doi.org/10.1002/9780470627242

[18] A'Hearn, B., Peracchi, F. and Vecchi, G. (2009) Height and the Normal Distribution: Evidence from Italian Military Data. *Demography*, **46**, 1-25. https://doi.org/10.1353/dem.0.0049

[19] Diverse Populations Collaborative Group (2005) Weight-Height Relationships and Body Mass Index: Some Observations from the Diverse Populations Collaboration. *The American Journal of Physical Anthropology*, **128**, 220-229. https://doi.org/10.1002/ajpa.20107

[20] Johnson, W., *et al.* (2020) Differences in the Relationship of Weight to Height, and Thus the Meaning of BMI According to Age, Sex, and Birth Year Cohort. *Annals of Human Biology*, **47**, 199-207. https://doi.org/10.1080/03014460.2020.1737731

[21] Sperrin, M., Marshall, A.D., Higgins, V., Renehan, A.G. and Buchan, I.E. (2015) Body Mass Index Relates Weight to Height Differently in Women and Older Adults: Serial Cross-Sectional Surveys in England (1992-2011). *Journal of Public Health*, **38**, 607-613. https://doi.org/10.1093/pubmed/fdv067

[22] Benn, R.T. (1971) Some Mathematical Properties of Weight-for-Height Indices Used as Measures of Adiposity. *Journal of Epidemiology & Community Health*, **25**, 42-50. https://doi.org/10.1136/jech.25.1.42

[23] Rohrer, F. (1921) Der Index der Körperfülle als Maß des Ernährungszustandes [The Index of Corpulence as a Measure of Nutritional Condition]. *Münchener Medizinische Wochenschrift*, **68**, 580-582.

[24] Henneberg, M., Hugg, J. and Townsend, E.J. (1989) Body Weight/Height Relationship: Exponential Solution. *American Journal of Human Biology*, **1**, 483-491.

https://doi.org/10.1002/ajhb.1310010412

[25] Cidras, M. (2015) Body Mass Exponential Index: An Age-Independent Anthropometric Nutritional Assessment. *Open Access Library Journal*, **2**, 1-8. https://doi.org/10.4236/oalib.1101943

[26] Trussell, J. and Bloom, D.E. (1979) A Model Distribution of Height or Weight at a Given Age. *Human Biology*, **51**, 523-536.

[27] Edwards, A.W.F. (1992) Likelihood. The Johns Hopkins University Press, Baltimore, 70-143.

[28] Kendall, M.G. and Stuart, A. (1963) The Advanced Theory of Statistics Vol. 1: Distribution Theory. Hafner, New York, 94-119, 228-236.

[29] Hotelling, H. (1953) New Light on the Correlation Coefficient and Its Transforms. *Journal of the Royal Statistical Society*: *Series B*, **15**, 193-232. https://doi.org/10.1111/j.2517-6161.1953.tb00135.x

[30] Mood, A.M., Graybill, F.A. and Boes, D.C. (1974) Introduction to the Theory of Statistics. 3rd Edition, McGraw-Hill, New York, 195-198, 233-236.

[31] Chou, Y. (1969) Statistical Analysis: With Business and Economic Applications. Holt, Rinehart, and Winston, New York, 308-323.

[32] Haldane, J.B.S. (1942) Moments of the Distributions of Powers and Products of Normal Variates. *Biometrika*, **32**, 226-242. https://doi.org/10.1093/biomet/32.3-4.226

[33] Arfken, G.B. and Weber, H.J. (2005) Mathematical Methods for Physicists. 6th Edition, Elsevier, New York, 83-87.

[34] Wikipedia (2022) Bootstrapping (Statistics). https://en.wikipedia.org/wiki/Bootstrapping_(statistics)

[35] Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, **7**, 1-26. https://doi.org/10.1214/aos/1176344552

[36] Harding, B., Tremblay, C. and Cousineau, D. (2014) Standard Errors: A Review and Evaluation of Standard Error Estimators Using Monte Carlo Simulations. *The Quantitative Methods for Psychology*, **10**, 107-123. https://doi.org/10.20982/tqmp.10.2.p107

[37] Moment (Mathematics) (2022) Wikipedia. https://en.wikipedia.org/wiki/Moment_(mathematics)

[38] Silverman, M.P., Strange, W. and Lipscombe, T.C. (2004) The Distribution of Composite Measurements: How to Be Certain of the Uncertainties in What We Measure. *American Journal of Physics*, **72**, 1068-1081. https://doi.org/10.1119/1.1738426

[39] Kendall, M.G. and Stuart, A. (1961) The Advanced Theory of Statistics Vol. 2: Inference and Relationship. Charles Griffin & Co., London, 1-8.

[40] Walker, J. (1996) HotBits: Genuine Random Numbers, Generated by Radioactive Decay. https://www.fourmilab.ch/hotbits

[41] Wikipedia (2022) Diehard Tests. https://en.wikipedia.org/wiki/Diehard_tests

[42] Maplesoft.com (2022) Overview of the RandomTools [MersenneTwister] Subpackage. https://www.maplesoft.com/support/help/maple/view.aspx

[43] Mapleprimes.com (2019) Are Maple's Pseudo Random Number Generators Good Generators? Post by mmcdara 3900. https://mapleprimes.com/posts/211598-Are-Maples-Pseudo-Random-Number-Generators

[44] Lapp, R.E. and Andrews, H.L. (1972) Nuclear Radiation Physics. Prentice-Hall, Englewood Cliffs, 36-40.

[45] Silverman, M.P., Strange, W., Silverman, C.R. and Lipscombe, T.C. (1999) Tests of Alpha-, Beta-, and Electron Capture Decays for Randomness. *Physics A*, **262**, 265-273. https://doi.org/10.1016/S0375-9601(99)00668-4

[46] Silverman, M.P. and Strange, W. (2009) Search for Correlated Fluctuations in the Decay of Na-22. *Europhysics Letters*, **87**, Article No. 32001. https://doi.org/10.1209/0295-5075/87/32001

[47] Silverman, M.P. (2015) Search for Non-Standard Radioactive Decay Based on Distribution of Activities. *Europhysics Letters*, **110**, Article No. 52001. https://doi.org/10.1209/0295-5075/110/52001

[48] Silverman, M.P. (2016) Search for Anomalies in the Decay of Radioactive Mn-54. *Europhysics Letters*, **114**, Article No. 62001. https://doi.org/10.1209/0295-5075/114/62001

[49] Miller, D.G. (1972) Radioactivity and Radiation Detection. Gordon and Breach, New York, 88-99.

[50] Foster, J., Kouris, K., Matthews, I.P. and Spyrou, N.M. (1983) Binomial vs Poisson Statistics in Radiation Studies. *Nuclear Instruments and Method*, **212**, 301-305. https://doi.org/10.1016/0167-5087(83)90706-8

Appendix—Glossary of Abbreviations

BMI—Body Mass Index
CDF—Cumulative Distribution Function
CF—Characteristic Function
d.o.f.—Degrees of Freedom
dCor—Distance Correlation
dCov—Distance Covariance
dVar—Distance Variance
ISNV—Independent Standard Normal Variable
PDF—Probability Density Function
RNG—Random Number Generator
RV—Random Variable