2022

Exact Statistical Distribution of the Body Mass Index (BMI): Analysis and Experimental Confirmation

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

Scientific Research Publishing

Open Journal of Statistics, 2022, 12, 324-356
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

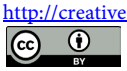
Exact Statistical Distribution of the Body Mass Index (BMI): Analysis and Experimental Confirmation

Mark P. Silverman[1*], Trevor C. Lipscombe[2]

[1]Department of Physics, Trinity College, Hartford, CT, USA
[2]The Catholic University of America, Washington, DC, USA
Email: *mark.silverman@trincoll.edu

How to cite this paper: 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

Received: April 5, 2022
Accepted: June 6, 2022
Published: June 9, 2022



Abstract

Body Mass Index (BMI), defined as the ratio of individual mass (in kilograms) to the square of the associated height (in meters), is one of the most widely discussed and utilized risk factors in medicine and public health, given the increasing obesity worldwide and its relation to metabolic disease. Statistically, BMI is a composite random variable, since human weight (converted to mass) and height are themselves random variables. Much effort over the years has gone into attempts to model or approximate the BMI distribution function. This paper derives the *mathematically exact* BMI probability density function (PDF), as well as the *exact* bivariate PDF for human weight and height. Taken together, weight and height are shown to be correlated bivariate lognormal variables whose marginal distributions are each lognormal in form. The mean and variance of each marginal distribution, together with the linear correlation coefficient of the two distributions, provide 5 nonadjustable parameters for a given population that uniquely determine the corresponding BMI distribution, which is also shown to be lognormal in form. The theoretical analysis is tested experimentally by gender against a large anthropometric data base, and found to predict with near perfection the profile of the empirical BMI distribution and, to great accuracy, individual statistics including mean, variance, skewness, kurtosis, and correlation. Beyond solving a longstanding statistical problem, the significance of these findings is that, with knowledge of the exact BMI distribution functions for diverse populations, medical and public health professionals can then make better informed statistical inferences regarding BMI and public health policies to reduce obesity.

Keywords

Body Mass Index, Obesity, Distribution of Weight, Distribution of Height,

Correlation of Weight and Height

1. Introduction

The body mass index (BMI) is a composite random variable defined by the relation [1]

$$B \equiv M/H^2 , \tag{1}$$

in which M is a person's mass in kg and H is the corresponding height in meters. It is to be recalled that a composite random variable comprises products and quotients (or a sum of products and quotients) of statistically distributed quantities [2]. As a readily obtainable quantitative measure of excess body fat, BMI is one of the most widely cited and discussed biomedical ratios employed by clinicians and epidemiologists [3]. Indeed, at the time of writing, BMI was the first item to come up when the phrase "most widely used biomedical index" was entered into the Google search engine. The reason for this is clear: obesity and its relation to metabolic disease are problems facing nearly all nations in both the developed and developing world [4].

Given its importance to individual medical treatments and public health policies, it is perhaps surprising that the statistical distribution of BMI from its inception to the present time has been uncertain and controversial. In this paper we show that weight (converted to mass) and height follow a correlated bivariate lognormal distribution, which leads to a uniquely specified lognormal distribution of BMI. A statistical test of our theoretical analysis by means of a large data base of individual mass, height, and BMI values provides strong evidence in support of our conclusion.[1]

1.1. Statistical Background of BMI

The concept of BMI was introduced as long ago as 1835 by Quetelet [5]. Initially assumed to be a normal (*i.e.* Gaussian) distribution by early developers of modern statistics, such as Galton and Pearson, the assumption was largely accepted by statisticians and scholars concerned with human growth throughout the 20th Century [5]. With the recognition that empirical BMI distributions appeared skewed to the right (*i.e.* to higher values), various non-symmetric distributions such as lognormal, gamma, beta, and power-law have been suggested, but none to our knowledge was rigorously demonstrated and tested. See, for example [6] [7] [8] [9].

The principal objective of this paper is to establish the distribution of BMI on

[1]In physics there is a difference between mass and weight. Excluding nuclear interactions, mass is an invariant; weight is the product of mass and gravitational acceleration and therefore depends on location and has different units than mass. In a medical context, weight is what is measured; statistically, it is mass that enters the BMI. Throughout this paper we refer to mass when analyzing the BMI distribution, but may speak of weight when referring to clinical studies, statistical sampling, data bases, and the like.

a more rigorous foundation and to test our findings experimentally against a large data base of personal weights and heights compiled in the 2012 Anthropometric Survey of US Army Personnel (ANSUR) [10]. Our analysis, discussed in the following sections, shows that, statistically, weight and height are correlated lognormal variables from which it rigorously follows that BMI is also a lognormal variable whose probability density function (PDF) is predictable from the 5 parameters that uniquely characterize the joint distribution of weight and height. The finding of lognormality is consistent with one of the authors (MPS) previous investigations [2] [11] of the distribution of composite random variables, which showed that such variables are ordinarily well represented by lognormal distributions irrespective of the distributions of the composite factors. A novel feature of BMI, however, not encountered in References [1] and [11] is the correlation of the individual factors of weight and height.

It has long been the practice in clinical medicine to use mean values of selected ratios, such as BMI to assess fat, LDL/HDL (low and high density lipoprotein)to assess cardiovascular risk, A/G (albumin and globulin) to assess liver function, BUN (blood urea nitrogen)/creatinine to assess kidney function, and many others. However, as emphasized by the authors in regard to statistical inferences [12] [11], the mean values alone may be uninformative and even misinformative. What is really required for valid interpretation and practical application of a biomedical index is its statistical distribution. By knowing the distribution of a random variable an analyst can determine with quantifiable confidence all population statistics (for comparison with empirical sample statistics) such as moments (mean, variance, skewness, kurtosis, etc.), cumulants, percentiles (such as median, quartiles, etc.), and, especially in the case of biomedical ratios, the cut-off values that determine degrees of health and risk.

In the analysis to follow, we calculate the exact distribution of BMI from its defining relation Equation (1), knowledge of the joint distribution of weight and height, and use of mathematical transformation relations for products and quotients of random variables [13] [14] [15]. The merit of this approach is that the form of the calculated distribution function is unique, apart from the empirical parameters that define the distributions of weight and height in a given population. Past attempts, such as cited above, to obtain mathematical expressions for the BMI distribution by curve-fitting data to assumed BMI profiles lack rigor and can actually lead to mathematically untenable results. For example, as cited above, statisticians have assumed throughout much of the 20th century that height, weight (or mass), and BMI all followed normal (Gaussian) distributions. It is mathematically demonstrable [2], however, that if mass and height are normal variables (which they are *not*), then the distribution of BMI *cannot possibly* also be normal, although the distribution profile may suggest such an appearance. Moreover, if mass and height are lognormal variables, then the distribution of BMI is *rigorously* lognormal too.

The virtue of having the exact (in contrast to an assumed or approximate) distribution is that it is expected to be valid for *all allowed values* of its parame-

ters and variables. Thus, since mass and height must take real, positive values, the BMI distribution must rigorously vanish at $B = 0$. A normal, or other approximate, distribution for BMI may appear to vanish at $B = 0$ if the mean of the distribution is sufficiently larger than the width (*i.e.* standard deviation); in other words, if the distribution is sharply defined. Rigorously, it does not vanish at $B = 0$. More problematic, however, is that a broad normal distribution can overlap the negative real axis leading to impossible values of BMI. Under such conditions, it may be thought that one could preserve normality simply by adopting a *truncated* normal distribution defined over the positive real axis. The outcome of truncation, although it may meet boundary conditions, will not fit data as well as the true distribution. Examples investigated by one of the authors (MPS) by means of the Principle of Maximum Entropy [16], have shown that a truncated Gaussian is *astronomically* less likely to be correct than the true distribution [11].

1.2. Significance of the BMI Distribution to Public Health

Before examining the statistical distribution of BMI, it is worth summarizing briefly how the distribution of BMI can influence the assessment of individual health and creation of public health policy.

The measurement of an individual BMI value requires only a person's weight and height. BMI therefore provides an inexpensive screening method for determining whether a person is underweight, healthy, overweight, or obese—the four general weight categories used by physicians and epidemiologists[2]. BMI is correlated with, and therefore seen as a proxy for, measurement of body fat [17], and is strongly correlated with metabolic disease [18]. Although there are other more accurate ways to measure body fat, such as bioimpedance analysis, dual-energy x-ray absorptiometry, computed tomography, and magnetic resonance imaging, such methods are expensive, not readily available to most patients or medical personnel, and require specially trained staff [17].

The statistical distribution of BMI provides the basis for setting the principal cut-off points that characterize various weight categories from severe thinness to severe obesity. Standard BMI cut-offs are independent of age and gender, although it is recognized that the same numerical value of BMI may correspond to different amounts of body fat in different populations, partly as a result of different body proportions [19]. More than 2 decades ago the World Health Organization (WHO) convened a working group of experts to study cut-offs in regard to the BMI of Asian populations, but the cut-offs remained largely unchanged [20] despite the conclusion that the proportion of Asians at high risk for type 2 diabetes and cardiovascular disease is substantial at BMIs *lower* than the existing WHO cut-off points for overweight. A recent report, based on a US

[2]The complete set of BMI classifications is more extensive than just four. Briefly, for illustrative purposes, according to the WHO a BMI ≥ 25 is considered overweight; ≥30 is considered obese; the range 18.5 - 24.9 is considered normal [1]. Nevertheless, there is much current discussion concerning the setting of BMI cutoff points.

population of adults and young adults reached a similar conclusion that a higher BMI only moderately increased the risks for diabetes among the healthy obese, and that unhealthy thin people were more likely than the aforementioned group to get diabetes [21]. Clearly, rigorous statistical distributions of BMI are needed for specific population groups. Otherwise, the matter of setting and interpreting BMI cut-off points will be remain controversial, not just in articles in the medical literature, but also in reports to the general public [22] [23].

In reference [22], questions were raised as to the accuracy and utility of using BMI to describe individual health. In our opinion, BMI was, and is, intended to be a statistical quantity. As such, it describes populations and not specific individuals. Nevertheless, given the exact distribution function specific to a well-defined demographic, cut-offs can be set more appropriately and less arbitrarily so as to be medically useful to clinicians in evaluating individual patients. To achieve this, a major first step would be to have an accurate BMI distribution function covering the full range of a sufficiently large and well-defined population of healthy individuals. We believe our analysis of the ANSUR data, which separately includes male and female members of the US military, provides such baseline information.

With regard to the establishment of public health policies to reduce adult obesity, knowledge of the exact BMI distribution can help resolve a debate over the optimal strategy for disease prevention. One approach, the "population strategy", proposed by Rose [24] [25] and widely adopted by epidemiologists, public health practitioners and policy makers, is to shift the distribution of a risk factor in a desired direction by applying interventions to an entire population [26]. An alternative approach, the "high-risk strategy", aims to lower the risk of disease within a population by detecting and treating the subgroup of people who manifest extreme values of the designated risk factor, and therefore appear to be at the highest risk.

Statistically, Rose and others found strong correlations between the mean value of a risk factor (e.g. BMI) and the prevalence of extreme values of that risk factor. In other words, in a specified population of people at risk for a particular disease over time, Rose expected the lower and middle sections of the distribution curve of the risk factor to move proportionally in the same direction as the high-end extreme section, thereby displacing the entire distribution curve to the right. He concluded from this, in regard to public health strategy, to implement a policy of intervention to *all* members of the population and not just those with risk factors in the upper tail of the distribution. The idea, as summarized in a review [26] of Rose's work, is that "More clinical cases result from small but widespread risks than large but rare risks."

Supporters of the "high-risk strategy" have pointed out problems to Rose's proposal. One such problem in regard to BMI in particular is the assumption of a Gaussian, or at least symmetric, distribution of the risk factor. This assumption, as we indicated in the previous section, is almost certainly invalid for any

realistic population distribution of human height and weight (mass). Speculations based on biological considerations have been made for a lognormal distribution [7], but such reasoning, while suggestive, is likewise not rigorous and leaves open the possibility of other skewed distribution functions.

Ultimately, which public health strategy is superior must be validated empirically. The efficacy of the "high-risk strategy" can be tested experimentally by randomized control trials. By contrast, it is more difficult to test the efficacy of the "population strategy". According to reference [26], the determination of whether a benefit results from lowering the risk of a whole population would require implementation and monitoring of lifestyle changes starting from birth and extending over decades.

Nevertheless, in order that an experimental test of strategy yield useful information, the results must be interpretable, and a valid interpretation requires the exact BMI probability density function. This function provides the most reliable statistical tool to study the evolution over time of the population statistics required for the "population strategy". Likewise, it helps the analyst decide quantitatively who falls within the category of high risk (*i.e.* proportion of a given population under the right tail of the distribution) as required for the "high risk strategy". From a broader perspective, the exact distribution function allows public health specialists to define meaningfully the cutoff points by which degrees of fatness and risk are classified.

Moreover, drawing upon the methodology of physics, we expect that, where direct experimental tests may be impractical to implement, the use of computer-based modeling can play a constructive role. Knowledge of the exact theoretical BMI distribution derived here, combined with well-designed mathematical models representing proposed public health interventions, can lead to insights and solutions in time intervals short compared to decades of observation.

2. Exact Distribution Function of BMI

The objective of this section is to derive the probability density function (PDF) of a random variable of the form of Equation (1), which we rewrite more generally as

$$Z = X/Y^2 = X/W \tag{2}$$

in which X and Y are arbitrary real-valued random variables and $W = Y^2$ with corresponding PDFs $p_X(x)$, $p_Y(y)$, $p_W(w)$. As a matter of standard notation, we represent a random variable by an upper case letter (e.g. X) and the realization or outcome of the variable (referred to as a variate) by the corresponding lower case letter (e.g. x).

Consider first the variable W, which must have non-negative variates. From the normalization criterion

$$\int_0^\infty p_W(w)\mathrm{d}w = \int_{-\infty}^\infty p_Y(y)\mathrm{d}y \tag{3}$$

or, equivalently, the differential transformation relation

$$p_W(w) = p_Y(y(w)) \Big/ \left|\frac{\mathrm{d}w}{\mathrm{d}y}\right| \tag{4}$$

with Jacobian $\left|\frac{\mathrm{d}w}{\mathrm{d}y}\right|$, one can derive the relation [13]

$$p_W(w) = \frac{1}{2\sqrt{w}}\left(p_Y\left(\sqrt{w}\right) + p_Y\left(-\sqrt{w}\right)\right). \tag{5}$$

An alternative and more versatile approach, which also leads to Equation (5), is to start with the defining transformation expressed by means of a Dirac delta function [27]

$$p_W(w) = \int_{-\infty}^{\infty} p_Y(y)\,\delta\left(y^2 - w\right)\mathrm{d}y\,. \tag{6}$$

The delta function, defined by the properties

$$\delta(x-y) \equiv \begin{cases} 0 & \text{if } x \neq y \\ \infty & \text{if } x = y \end{cases} \tag{7}$$

with unit area

$$\int_{-\infty}^{\infty} \delta(x-y)\mathrm{d}x = 1\,, \tag{8}$$

is not actually a function, but what mathematicians refer to as a δ-distribution and physicists as a unit impulse. From its definition follows useful operational relations

$$\int_{-\infty}^{\infty} f(x)\,\delta(x-x_0)\mathrm{d}x = f(x_0) \tag{9}$$

$$\delta(ax) = \frac{1}{|a|}\delta(x)\,, \tag{10}$$

$$\delta\left(x^2 - a^2\right) = \frac{1}{2|a|}\left(\delta(x-a) + \delta(x+a)\right) \tag{11}$$

$$\delta(g(x)) = \sum_i \frac{\delta(x-x_i)}{\left|\mathrm{d}g/\mathrm{d}x\right|_{x=x_i}} \tag{12}$$

where a is a constant and $g(x)$ a continuous real-valued function with zero points at x_i, *i.e.* $g(x_i) = 0$. As seen from Equation (9), the delta function serves as a filtering operation in integration. It can be represented in numerous ways by a limiting process of which one commonly used form is the Fourier transform of unity

$$\delta(x-y) = \lim_{K\to\infty} \int_{-K}^{K} \mathrm{e}^{ik(x-y)}\mathrm{d}k\,. \tag{13}$$

Consider next the quotient $Z = X/W = X/Y^2$ for independent variables X and Y. Starting with the defining transformation

$$p_Z(z) = \int_{-\infty}^{\infty} \mathrm{d}x \int_0^{\infty} \mathrm{d}w\, p_X(x)\, p_W(w)\, \delta\left(z - \frac{x}{w}\right) \tag{14}$$

and employing relations (9)-(12) reduces integral (14) to the form

$$p_Z(z) = \int_0^{\infty} y^2 p_X\left(zy^2\right)\left(p_Y(y) + p_Y(-y)\right)\mathrm{d}y\,. \tag{15}$$

Upon identification of X with mass and Y with height, Equation (15) is the *exact* distribution function of the random variable B representing body mass index under the condition that mass and height are statistically uncorrelated. We examine the case of correlated mass and height in Section 2.3.

2.1. Special Case: Independent Normal Factors

Over much of the period of modern statistics human attributes such as height and weight have been assumed or approximated to follow a Gaussian distribution. Justification for this may be attributed in part to empirical inferences drawn from coarse-graded statistical sampling, theoretical inferences based on the Central Limit Theorem, and a need for mathematical convenience [28]. Statisticians were certainly aware, however, that the tails of a Gaussian did not fit observed frequency data closely [29] [30], but this problem was generally regarded as minor since the number of events were few compared with the bulk of the observed frequency distribution. With regard to BMI, however, the tail of the distribution is important since it represents the subgroup of people with extreme risk factors. Nevertheless, because normal distributions serve as a kind of baseline model in the statistics of public health, we examine the case of independent normally distributed weight (mass) X and height Y,

$$\begin{aligned} X &= N_X\left(m_1, s_1^2\right) \\ Y &= N_Y\left(m_2, s_2^2\right) \end{aligned} \tag{16}$$

represented statistically by the symbol $N\left(m, s^2\right)$ in which m is the mean of the variable, s^2 is the variance, and the PDF of a normal distribution (indicated by superscript N)takes the general form

$$p_X^{(N)}(x) = \frac{1}{\sqrt{2\pi s^2}} \mathrm{e}^{-(x-m)^2/2s^2}\,. \tag{17}$$

Substitution into Equation (15) of PDF (17) with the parameters of distributions (16) leads to the explicit function

$$p_Z^{(N,N)}(z) = \frac{1}{2\pi s_1 s_2} \int_0^{\infty} y^2 \mathrm{e}^{-\left(zy^2 - m_1\right)^2/2s_1^2} \left(\mathrm{e}^{-(y-m_2)^2/2s_2^2} + \mathrm{e}^{-(y+m_2)^2/2s_2^2}\right)\mathrm{d}y\,. \tag{18}$$

where the superscript (N,N) signifies that both component factors (mass and height) are normally distributed. It is clear from the form of Equation (18), in which the leading term in the exponent is z to the 4th power (rather than quadratic), that normal distributions of mass and height, as expressed in relations (16), result in a *non*-normal and *non*-symmetrical distribution of body mass in-

dex.

To our knowledge, the integral (18) cannot be performed analytically, but must be evaluated numerically. A plot of $p_Z^{(N,N)}(z)$ (solid curve) as a function of z for a hypothetical sample set of parameters is shown in Figure 1. The profile is skewed to the right and appears very much like the lognormal profile (dashed curve), superposed for comparison. As seen in the figure, the two profiles are distinguishable, but in close agreement, especially in the vicinity of the maximum, the origin, and along the right tail. If the range of the figure were extended to show the tail out to $z = 200$, the two profiles would appear to overlap apart from a slightly lower maximum value of the lognormal distribution.

The lognormal distribution of BMI shown in Figure 1, discussed more fully in the following sections, is the profile that would result if individual distributions of mass and height were both lognormal with parameters corresponding to the parameters of the normal distributions in the example. The PDF of the variable $Z = X/Y^2$ in which X and Y are lognormal variables takes the general form

$$p_Z^{(\Lambda,\Lambda)}(z) = \frac{1}{\sqrt{2\pi s^2}\, z} \mathrm{e}^{-\frac{(\ln(z)-m)^2}{2s^2}} \tag{19}$$

to be derived in Section 2.3. The superscript (Λ,Λ) signifies that both component factors are lognormal. Determination of the lognormal parameters corresponding to the parameters of the normal distributions that form Figure 1 is explained in the following section.

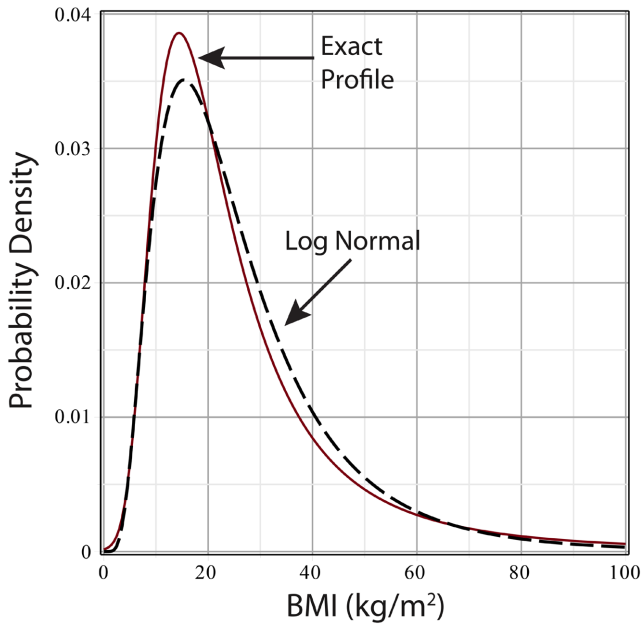


Figure 1. Exact probability density (solid maroon curve) for $Z = X/Y^2$ for mass $X = N(70,20)$ and height $Y = N(1.8,0.5)$. Superposed is the corresponding lognormal density (dashed black curve) for $Z = \Lambda(3.1080, 0.6130^2)$. The relation between the normal and lognormal parameters is explained in Section 2.2. It is to be noted that in actuality human weight and height are *not* distributed normally.

The exact PDF (18) and the corresponding lognormal PDF (19) yield, respectively, the following means, dispersions (standard deviation about the mean), and asymmetries (skewness, defined in the next section)

$$\begin{array}{lll} \text{Mean} & \mu_Z^{(N,N)} = 27.25 & \mu_Z^{(\Lambda,\Lambda)} = 27.00 \\ \text{Std Dev} & \sigma_Z^{(N,N)} = 21.93 & \sigma_Z^{(\Lambda,\Lambda)} = 18.23 \\ \text{Skewness} & Sk_Z^{(N,N)} = 2.68 & Sk_Z^{(\Lambda,\Lambda)} = 2.33 \end{array} \tag{20}$$

which are seen to be numerically close for the exact and lognormal distributions of BMI.

The statistics exhibited in relation (20) raise a cautionary issue in regard to skewed distribution functions. Ordinarily—*i.e.* primarily for symmetric distributions—the standard deviation is interpreted as a measure of the uncertainty of the mean, which, itself, is usually adopted in statistical physics and medicine as the experimental value of a distributed quantity. However, as seen in **Figure 1**, the modes (*i.e.* maxima), in contrast to the means, of the two profiles are actually fairly narrowly located and serve better than the mean for purposes of monitoring the evolution of the BMI distribution in a specified population over time. The large dispersions about the means are due to the long high-end tails. The skewness of a distribution, which is proportional to the 3rd central moment, provides a quantitative measure of the asymmetry about the mean, and therefore a measure of the fraction of a population at greatest risk of metabolic disease.

In summary, for purposes of defining appropriate cutoff points for the various weight categories and demographics and to investigate evolving trends in BMI within a population, a lognormal distribution would serve equally satisfactory to the exact BMI distribution derived on the assumption of normally distributed weight and height. Our analysis indicates, however, that this assumption is *not* valid, and that the true distributions of weight and height are, themselves, lognormal, from which it follows that a lognormal BMI distribution is not an approximation, but rigorously exact. We discuss this in the following section.

2.2. Special Case: Independent Lognormal Factors

If the natural logarithm of a set of variates $\{x_i\}$, represented by the random variable X, gives rise to a normal distribution, represented by the variable Y, then X is said to be a lognormal random variable. The relation is expressed symbolically as

$$Y = \ln(X) = N_Y\left(m, s^2\right) \;\Rightarrow\; X = \exp(Y) = \Lambda_X\left(m, s^2\right). \tag{21}$$

From the transformation relations (21) and normal PDF (17), there follows the lognormal PDF

$$p_X^{(\Lambda)}(x) = \frac{1}{\sqrt{2\pi s^2}} \frac{\mathrm{e}^{-(\ln(x)-m)^2/2s^2}}{x}. \tag{22}$$

Note that the parameters defining the lognormal distribution X are the mean and variance of the *normal* variable Y. In other words, m and s^2 are *not* mo-

ments of the lognormal distribution. The r^{th} order moments $M_0(r) \equiv \langle X^r \rangle$, $r = 1, 2, \cdots$, of a lognormal distribution can be calculated straightforwardly as expectation values by using PDF (22)

$$\langle X^r \rangle = \int_0^\infty x^r p_X^{(\Lambda)}(x) \mathrm{d}x . \tag{23}$$

However, it is simpler to calculate $M_0(r)$ from the moment generating function [13]

$$g_Y(t) \equiv \langle \exp(Yt) \rangle = \mathrm{e}^{mt + \frac{1}{2}s^2 t^2} = \langle X^t \rangle \tag{24}$$

of the normal distribution Y by use of relation (21) and substitution of the discrete index r for the continuous dummy variable t. This leads to

$$M_0(r) = \mathrm{e}^{mr + \frac{1}{2}s^2 r^2} . \tag{25}$$

The first few moments $M_0(r)$ and the principal combination statistics

$$\text{Variance} \qquad \sigma_X^2 \equiv \langle (X - \mu)^2 \rangle \tag{26}$$

$$\text{Skewness} \qquad Sk \equiv \langle (X - \mu)^3 \rangle \big/ \sigma_X^3 \tag{27}$$

$$\text{Kurtosis} \qquad K_X \equiv \langle (X - \mu)^4 \rangle \big/ \sigma_X^4 \tag{28}$$

of the lognormal distribution are summarized in Table 1. The subscript 0 in $M_0(r)$ signifies that the moments are taken with respect to the origin. The preceding combination statistics are central moments, designated $M_\mu(r)$ in Table 1, where the subscript μ signifies that the moments are taken with respect to the lognormal mean $\mu \equiv M_0(1)$

$$M_\mu(r) = \int_0^\infty p_X(x)(x - \mu)^r \, \mathrm{d}x = \sum_{j=0}^{r} (-1)^{r-j} C(r, j) \mu^{r-j} M_0(j) . \tag{29}$$

The symbol

$$C(r, j) \equiv \frac{r!}{j!(r - j)!} \tag{30}$$

is a binomial coefficient.

The mean μ and variance σ^2 of the lognormal variable X is given in terms of the mean m and variance s^2 of the normal variable Y by the following relations from Table 1.

$$\begin{aligned} \mu &= \mathrm{e}^{m + \frac{1}{2}s^2} \\ \sigma^2 &= \mathrm{e}^{2m}\left(\mathrm{e}^{2s^2} - \mathrm{e}^{s^2}\right) \end{aligned} \tag{31}$$

from which follow the inverse relations

$$m = \ln\left(\frac{\mu^2}{\sqrt{\mu^2 + \sigma^2}}\right) \tag{32}$$

Table 1. Moments and variances of the log-normal distribution.

Order r	Moment $M_0(r)$	Central Moment: $M_\mu(r)$
1	$e^{m+\frac{1}{2}s^2}$	0
2	e^{2m+2s^2}	$e^{2m}\left(e^{2s^2}-e^{s^2}\right)$
3	$e^{3m+\frac{9}{2}s^2}$	$e^{3m}\left(e^{\frac{9}{2}s^2}-3e^{\frac{5}{2}s^2}+2e^{\frac{3}{2}s^2}\right)$
4	e^{4m+8s^2}	$e^{4m}\left(e^{8s^2}-4e^{5s^2}+6e^{3s^2}-3e^{2s^2}\right)$
5	$e^{5m+\frac{25}{2}s^2}$	$e^{5m}\left(e^{\frac{25}{2}s^2}-5e^{\frac{17}{2}s^2}+10e^{\frac{11}{2}s^2}-10e^{\frac{7}{2}s^2}+4e^{\frac{5}{2}s^2}\right)$
6	e^{6m+18s^2}	$e^{6m}\left(e^{18s^2}-6e^{13s^2}+15e^{9s^2}-20e^{6s^2}+15e^{4s^2}-5e^{3s^2}\right)$
7	$e^{7m+\frac{49}{2}s^2}$	$e^{7m}\left(e^{\frac{49}{2}s^2}-7e^{\frac{37}{2}s^2}+21e^{\frac{27}{2}s^2}-35e^{\frac{19}{2}s^2}+35e^{\frac{13}{2}s^2}-21e^{\frac{9}{2}s^2}+6e^{\frac{7}{2}s^2}\right)$
8	e^{8m+32s^2}	$e^{8m}\left(e^{32s^2}-8e^{25s^2}+28e^{19s^2}-56e^{14s^2}+70e^{10s^2}-56e^{7s^2}+28e^{5s^2}-7e^{4s^2}\right)$
Statistic	**Symbol**	**Expression**
Standard Deviation	σ	$e^m\sqrt{\left(e^{2s^2}-e^{s^2}\right)}$
Skewness	Sk	$\left(e^{3s^2}-3e^{s^2}+2\right)\Big/\left(e^{s^2}-1\right)^{\frac{3}{2}}$
Kurtosis	K	$e^{4s^2}+2e^{3s^2}+3e^{2s^2}-3$
Variance of Skewness	$Var(Sk)$	$\frac{Var\left(M_\mu(3)\right)}{S^6}+\frac{3M_\mu(3)^2Var\left(M_\mu(2)\right)}{4S^{13}}$
Variance of Kurtosis	$Var(K)$	$\frac{Var\left(M_\mu(4)\right)}{S^8}+\frac{2M_\mu(4)^2Var\left(M_\mu(2)\right)}{S^{16}}$

$$s^2=\ln\left(\frac{\mu^2+\sigma^2}{\mu^2}\right). \tag{33}$$

Relations (31), (32), (33) will be applied shortly to the BMI distribution in Figure 1.

Consider next the case of independent lognormal factors for mass and height respectively

$$\begin{aligned} X&=\Lambda_X\left(m_1,s_1^2\right)\\ Y&=\Lambda_Y\left(m_2,s_2^2\right)\end{aligned} \tag{34}$$

resulting in the BMI[3]

[3]We use symbols X and Y at various points in the paper to represent different types of random variables in different examples. This should pose no difficulty because in each case the distribution of each variable is precisely defined at the outset. We believe it is easier for the reader to keep track of just two symbols in a discussion than burden this paper with a different symbol each time a variable is used in an example.

$$Z = X/Y^2 . \tag{35}$$

To find the distribution of Z, take the natural logarithm of both sides of expression (35) to obtain

$$\begin{aligned} \ln(Z) &= \ln(X) - 2\ln(Y) \\ &= N\left(m_1, s_1^2\right) - 2N\left(m_2, s_2^2\right) \\ &= N\left(m_1 - 2m_2, s_1^2 + 4s_2^2\right) \end{aligned} \tag{36}$$

where the first equality of the second line of Equation (36) follows from the definition of a lognormal variable, and the second equality is the result of combining two independent normal distributions, which follows from the equivalence relation [13],

$$N\left(m, s^2\right) = m + sN(0,1) . \tag{37}$$

Thus, since the log of Z is a normal variable, then Z must be a lognormal variable $Z = \Lambda_Z\left(m, s^2\right)$ with parameters

$$\begin{aligned} m &= m_1 - 2m_2 \\ s^2 &= s_1^2 + 4s_2^2 \end{aligned} \tag{38}$$

From relations (31), the parameters (38) correspond to a mean BMI of

$$\mu = M_0(1) = \mathrm{e}^{(m_1 - 2m_2) + \frac{1}{2}\left(s_1^2 + 4s_2^2\right)} \tag{39}$$

with standard deviation

$$\sigma = \sqrt{M_0(2) - M_0(1)^2} = \mathrm{e}^{(m_1 - 2m_2) + \frac{1}{2}\left(s_1^2 + 4s_2^2\right)} \sqrt{\mathrm{e}^{2\left(s_1^2 + 4s_2^2\right)} - \mathrm{e}^{\left(s_1^2 + 4s_2^2\right)}} . \tag{40}$$

In the example illustrated in **Figure 1**, a hypothetical sample population was characterized statistically by mass $\mu_M \pm \sigma_M = 70 \pm 20$ kg and height $\mu_H \pm \sigma_H = 1.8 \pm 0.5$ m. If mass and height independently follow the respective lognormal distributions $\Lambda_X\left(m_1, s_1^2\right)$ and $\Lambda_Y\left(m_2, s_2^2\right)$, the four parameters of the distributions are rigorously determined from Equations (31)-(33) (to four decimal places)

$$\begin{aligned} m_1 &= 4.2093, \quad & s_1 &= 0.2801 \\ m_2 &= 0.5506, \quad & s_2 &= 0.2726 \end{aligned} \tag{41}$$

The corresponding parameters of the BMI distribution $\Lambda_Z\left(m, s^2\right)$ in **Figure 1** (dashed curve) are then given by Equation (38)

$$m = 3.1080, \quad s = 0.6130 . \tag{42}$$

From the inverse relations (39) and (40) one calculates the mean and standard deviation of the BMI distribution to be

$$\mu_{BMI} = 27.0018, \quad \sigma_{BMI} = 18.2349 , \tag{43}$$

which agree with the corresponding values in Equation (20) obtained by integration over the PDF as in Equation (23).

2.3. Special Case: Correlated Lognormal Factors

The exact BMI distribution expressed by Equation (15) contains in the integrand

products of the PDFs of the variables X and Y. However, if the weight of an individual is influenced by his/her height (or *vice versa*), then the *joint* distribution function of mass and height, expressed as $p_{XY}(x,y)$, does not factorize into separate functions of x and y. In that case, the antecedent Equation (14) with $W = Y^2$ takes the form

$$p_Z(z) = \int_0^\infty \mathrm{d}x \int_0^\infty \mathrm{d}w\, p_{XW}(x,w)\,\delta\left(z - \frac{x}{w}\right) \tag{44}$$

leading to the result

$$p_Z(z) = \int_0^\infty y^2 \left(p_{XY}\left(zy^2, y\right) + p_{XY}\left(zy^2, -y\right)\right)\mathrm{d}y\,. \tag{45}$$

In the following section we provide strong evidence that height and weight *are* significantly correlated and that the marginal distributions of both variables are lognormal in form. The joint distribution function of bivariate *lognormal* variables is derivable from the distribution function of bivariate *normal* variables [31]

$$\begin{aligned} p_{Y_1Y_2}^{(N)}(y_1,y_2) &= \frac{1}{2\pi s_1 s_2\sqrt{1-r^2}}\,\mathrm{e}^{-q_Y/2} \\ q_Y &= \frac{1}{1-r^2}\left[\left(\frac{y_1-m_1}{s_1}\right)^2 - 2r\left(\frac{y_1-m_1}{s_1}\right)\left(\frac{y_2-m_2}{s_2}\right) + \left(\frac{y_2-m_2}{s_2}\right)^2\right] \end{aligned} \tag{46}$$

In the preceding equation, m_1, s_1 are the mean and standard deviation of a normal variable Y_1, and likewise m_2, s_2 are the mean and standard deviation of a normal variable Y_2. The Pearson correlation coefficient r is defined as the expectation value [31]

$$r \equiv \frac{\left\langle (Y_1 - m_1)(Y_2 - m_2)\right\rangle}{s_1 s_2} = \frac{\mathrm{cov}(Y_1, Y_2)}{s_1 s_2}, \tag{47}$$

which falls within the range $1 \geq r \geq -1$. The expectation value in the numerator of Equation (47) is the covariance. A correlation coefficient $r = 1$ signifies that the two variables are perfectly correlated linearly; likewise, $r = -1$ signifies perfect linear anticorrelation. An arbitrary value of r within the stated range is interpreted to mean that r^2 is the fraction of the variance of one variable attributable to the other [32].

The probability density function $p_R(r)$ of the Pearson r is a complicated mathematical expression involving gamma functions and a hypergeometric function of the type ${}_2F_1$. The exact form of the PDF and resulting statistical moments can be found in Ref. [33]. Of particular utility in this paper is the standard deviation σ_r and standard error (SE)

$$SE_r \equiv \frac{\sigma_r}{\sqrt{n}} = \frac{1-r^2}{\sqrt{n}} \tag{48}$$

truncated at the first term of an expansion in inverse powers of the sample size n. Plots of $p_R(r)$ for different mean values $\langle r\rangle$ and two sample sizes n are

displayed in Figure 2. The profiles are strongly skewed to the left for small sample size and rapidly approach Gaussian form as n increases. For the ANSUR data used in this paper, $n > 1000$ and the exact profile of $p_R(r)$ is indistinguishable from a normal distribution about $\langle r \rangle$ with width SE_r given by (48).

If the normally distributed variates of Y_1 and Y_2 are obtained by taking the natural logarithm of the variates of X_1 and X_2, then X_1 and X_2 are log-normal variables. In analogy to Equation (4), the transformation of PDF (46) to a PDF of X_1 and X_2 is implemented as follows

$$p_{X_1X_2}^{(\Lambda,\Lambda)}(x_1,x_2) = p_{Y_1Y_2}^{(N,N)}\left(y_1(x_1),y_2(x_2)\right)\left|\frac{dy_1}{dx_1}\frac{dy_2}{dx_2}\right| \tag{49}$$

where $y(x) = \ln(x)$ and leads to

$$\begin{aligned} p_{X_1X_2}^{(\Lambda,\Lambda)}(x_1,x_2) &= \frac{1}{2\pi s_1 s_2\sqrt{1-r^2}}\frac{e^{-q_X/2}}{x_1x_2} \\ q_X &= \frac{1}{1-r^2}\left[\left(\frac{\ln(x_1)-m_1}{s_1}\right)^2 - 2r\left(\frac{\ln(x_1)-m_1}{s_1}\right)\left(\frac{\ln(x_2)-m_2}{s_2}\right)+\left(\frac{\ln(x_2)-m_2}{s_2}\right)^2\right] \end{aligned} \tag{50}$$

which generalizes Equation (22).

The marginal distribution of one variable is obtained by integrating the PDF (50) over the other variable as follows

$$\begin{aligned} \int_0^\infty p_{X_1X_2}^{(\Lambda,\Lambda)}(x_1,x_2)\,dx_2 &= p_{X_1}^{(\Lambda)}(x_1) \\ \int_0^\infty p_{X_1X_2}^{(\Lambda,\Lambda)}(x_1,x_2)\,dx_1 &= p_{X_2}^{(\Lambda)}(x_2) \end{aligned} \tag{51}$$

As one would expect, the correlation coefficient r vanishes from the marginal distributions, since both variables must be present if there is to be a correlation between them.

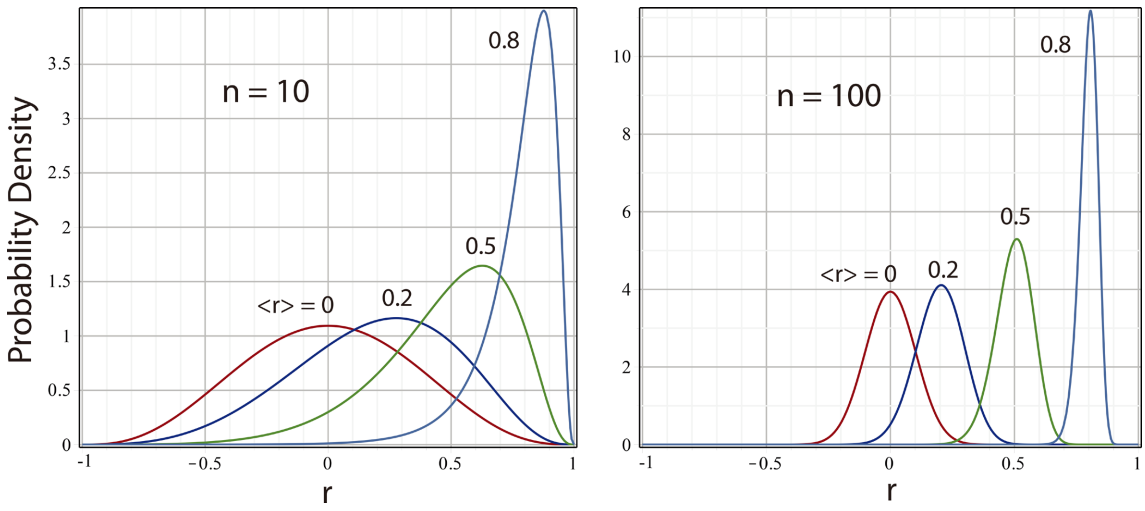


Figure 2. Probability density of Pearson r coefficient for different values of the mean $\langle r \rangle$ and sample sizes $n = 10$ (left panel) and $n = 100$ (right panel). The PDF rapidly approaches Gaussian form in the limit of increasing sample size.

It is to be borne in mind that the Pearson coefficient r is a measure of the correlation between *normal* variables Y_1 and Y_2. The Pearson coefficient of the *lognormal* variables X_1 and X_2, which represent respectively mass and height in the context of BMI, is obtained from the relation corresponding to (47)

$$\rho = \frac{\operatorname{cov}\left(\left(X_1 - \mu_1\right)\left(X_2 - \mu_2\right)\right)}{\sigma_1 \sigma_2} \tag{52}$$

which can be reduced to

$$\rho = \frac{\left\langle \left(X_1 - \mu_1\right)\left(X_2 - \mu_2\right)\right\rangle}{\sigma_1 \sigma_2} = \frac{\left\langle X_1 X_2 \right\rangle - \mu_1 \mu_2}{\sigma_1 \sigma_2}. \tag{53}$$

Equation (53) requires the integral

$$\left\langle X_1 X_2 \right\rangle = \int_0^\infty x_1 \mathrm{d}x_1 \int_0^\infty x_2 p_{X_1 X_2}\left(x_1, x_2\right) \mathrm{d}x_2 = \mu_1 \mu_2 \mathrm{e}^{r s_1 s_2} \tag{54}$$

where mean value μ of a lognormal variable is given by Equation (31). From Equations (54), (53), and (31), it follows that the correlation coefficient ρ of a bivariate lognormal distribution takes the form

$$\rho = \frac{\mathrm{e}^{r s_1 s_2} - 1}{\sqrt{\left(\mathrm{e}^{s_1^2} - 1\right)\left(\mathrm{e}^{s_2^2} - 1\right)}}. \tag{55}$$

It is worth noting that the moments, including all correlation statistics of a bivariate, or more generally a multivariate, distribution can in principle be obtained from a moment generating function [13] without having to perform integrals like the one in Equation (54), which can be difficult. This method, however, lies outside the scope of this paper. Nevertheless, integrations over the bivariate PDF (50) can be greatly simplified by transforming from the space of $\left(x_1, x_2\right)$ back to the space of $\left(y_1, y_2\right)$ and then transforming to variables $\left(u, v\right)$ defined by

$$\begin{aligned} u &= \frac{y_1 - m_1}{\sqrt{1 - r^2}\, s_1} \\ v &= \frac{y_2 - m_2}{\sqrt{1 - r^2}\, s_2} \end{aligned} \tag{56}$$

which generates the probability density

$$f\left(u, v\right) = \frac{\sqrt{1 - r^2}}{2\pi} \mathrm{e}^{-\frac{1}{2}\left(u^2 - 2ruv + v^2\right)}. \tag{57}$$

The range of variables u, v is $\left(-\infty, \infty\right)$. To calculate joint expectations of powers of X_1 and X_2, substitute

$$\begin{aligned} x_1 &= \mathrm{e}^{y_1} = \mathrm{e}^{\sqrt{1 - r^2} s_1 u + m_1} \\ x_2 &= \mathrm{e}^{y_2} = \mathrm{e}^{\sqrt{1 - r^2} s_2 v + m_2} \end{aligned} \tag{58}$$

in the integral with PDF $f\left(u, v\right)$.

Even with the preceding transformations to facilitate calculation, we have

been unable to derive in closed form an expression for the variance of Equation (55). We approximate, therefore, the variance of ρ by using error propagation theory [34]

$$\sigma_\rho^2 = \left(\frac{\partial \rho}{\partial r}\right)^2 \sigma_r^2 + \left(\frac{\partial \rho}{\partial s_1^2}\right)^2 \sigma_{s_1}^2 + \left(\frac{\partial \rho}{\partial s_2^2}\right)^2 \sigma_{s_2}^2 \tag{59}$$

in which σ_r^2 is given by Equation (48), and the variance of the variance s^2 of a normal random variable $N\left(m, s^2\right)$ is known to be [13]

$$\sigma_{s^2}^2 = 2s^4 . \tag{60}$$

Standard errors are obtained by dividing the variances σ_r^2, $\sigma_{s_1^2}^2$, $\sigma_{s_2^2}^2$ by the sample size *n*. The analytical evaluation of Equation (59) leads to a long, and not particularly illuminating expression and will not be given explicitly, since, when its evaluation is needed later, both the partial derivatives and numerical substitutions are carried out by computer.

Calculation of the probability density function of the ratio $Z = X_1 / X_2^2$ using PDF (50) with lognormal factors $X_1 = \Lambda_{X_1}\left(m_1, s_1^2\right)$ for mass and $X_2 = \Lambda_{X_2}\left(m_2, s_2^2\right)$ for height proceeds most readily from the defining transformation

$$\begin{aligned} p_Z(z) &= \int_0^\infty \int_0^\infty p_{X_1X_2}(x_1, x_2)\, \delta\left(z - \frac{x_1}{x_2^2}\right) \mathrm{d}x_1 \mathrm{d}x_2 \\ &= \int_0^\infty x_2^2\, p_{X_1X_2}\left(zx_2^2, x_2\right) \mathrm{d}x_2 \end{aligned} \tag{61}$$

where the second line of relation (61) results immediately from property (10) of the delta function. The remaining integration can be performed by transforming to the integration variable $y_2 = \ln\left(x_2\right)$ and leads to the exact PDF

$$p_Z^{(\Lambda)}(z) = \frac{\mathrm{e}^{-\frac{\left(\ln(z) - (m_1 - 2m_2)\right)^2}{2\left(s_1^2 + 4s_2^2 - 4rs_1s_2\right)}}}{\sqrt{2\pi\left(s_1^2 + 4s_2^2 - 4rs_1s_2\right)z}} \tag{62}$$

for BMI of a population with correlated weight and height.

From the form of PDF (62), it is seen that the variable *Z* is exactly lognormal

$$Z = \Lambda_Z\left(m, s^2\right), \tag{63}$$

with parameters

$$\begin{aligned} m &= m_1 - 2m_2 \\ s^2 &= s_1^2 + 4s_2^2 - 4rs_1s_2 \end{aligned} \tag{64}$$

Comparison with Equation (38) shows that the mean *m* is the same as for independent lognormal factors, but the variance s^2 is a function of the correlation coefficient *r*. The influence of correlation on the probability density (and therefore also on statistical moments)can be quite strong, as illustrated in **Figure 3** which shows plots of PDF (62) as a function of the BMI variate *z* for values of *r* ranging from −1 to +1 in intervals of 0.25. The plots are color coded such that

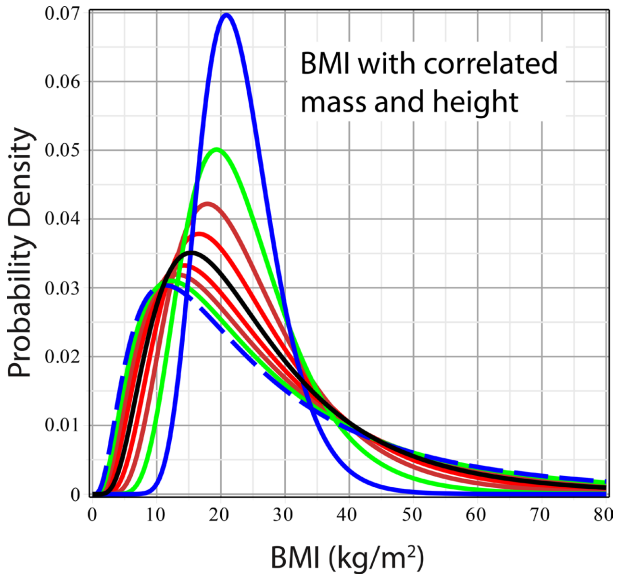


Figure 3. Exact BMI distribution for lognormally distributed correlated mass and height. The correlation coefficient $r = +1$ (solid blue), -1 (dashed blue), 0 (solid black) and varies from minimum to maximum in increments of 0.25. Positive correlation leads to narrower profiles. The parameters of the marginal mass and height distributions are the same as for **Figure 1**.

profiles of the same $|r|$ have the same color, but are distinguished by their widths ranging from a maximum for $r = -1$ (dashed blue curve) to a minimum of $r = 1$ (solid blue curve). The solid black profile corresponds to uncorrelated weight and height, $r = 0$. As shown in the figure, increasing the correlation of weight and height displaces the maximum of the BMI distribution to the right and narrows the spread. For perfect linear correlation $r = 1$, the variance takes its minimum value, $s^2\big|_{\min} = (s_1 - 2s_2)^2$, which, as expected, can never be negative. As a corollary of the narrower spread, the tail of the BMI distribution with positive correlation drops off more rapidly than if weight and height were uncorrelated or anticorrelated.

BMI population statistics, of which the most important are the mean, dispersion about the mean (standard deviation), and asymmetry about the mean (skewness)

$$\mu_Z \equiv \langle Z \rangle = \exp\left((m_1 - 2m_2) + \frac{1}{2}\left(s_1^2 + 4s_2^2 - 4rs_1 s_2 \right) \right) \tag{65}$$

$$\sigma_Z \equiv \sqrt{\langle Z^2 \rangle - \langle Z \rangle^2} = \mathrm{e}^{(m_1 - 2m_2)} \sqrt{\mathrm{e}^{\left(2s_1^2 + 8s_2^2 - 8rs_1 s_2\right)} - \mathrm{e}^{\left(s_1^2 + 4s_2^2 - 4rs_1 s_2\right)}} \tag{66}$$

$$Sk_Z \equiv \frac{\left\langle (Z - \mu)^3 \right\rangle}{\sigma_Z^3} = \frac{\mathrm{e}^{3\left(s_1^2 + 4s_2^2 - 4rs_1 s_2\right)} - 3\mathrm{e}^{s_1^2 + 4s_2^2 - 4rs_1 s_2} + 2}{\left(\mathrm{e}^{s_1^2 + 4s_2^2 - 4rs_1 s_2} - 1\right)^{3/2}} \tag{67}$$

are also markedly affected by the correlation of weight and height, as plotted in **Figure 4** as a function of correlation coefficient r. As shown in the figure, the higher the correlation, the lower are the BMI mean, standard deviation, and skewness.

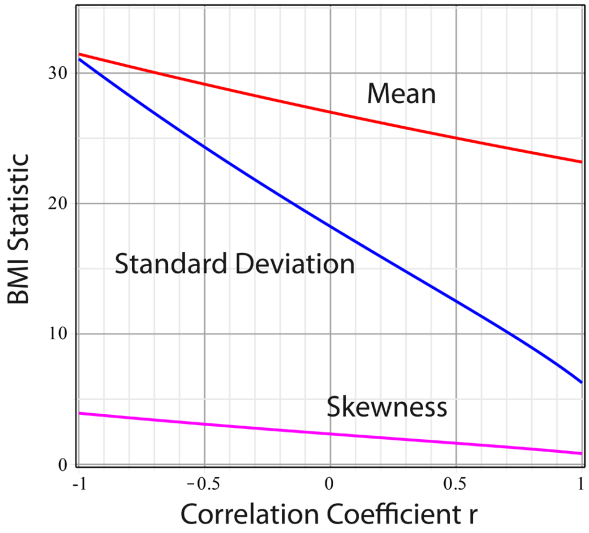


Figure 4. Variation of lognormal BMI mean (red), standard deviation (blue), and skewness (magenta) with Pearson correlation coefficient *r*. The parameters of the marginal mass and height distributions are the same as for Figure 1.

3. Statistical Analysis of the ANSUR Data

The Anthropometric Survey of U.S. Army Personnel (ANSUR), conducted in 2012 and reported in 2014 [10], was undertaken by the Natick Soldier Research, Development and Engineering Center (NSRDC) in Natick, Massachusetts to obtain an extensive body of data from comparably measured individuals representative of the "Total Army" of active-duty personnel. The motivation of the survey was to obtain accurate data by which the Army could make appropriate decisions regarding clothing, protective equipment, workspaces, and other size-dependent, work-related matters.

In keeping with this need, the survey measured 93 dimensions directly and 41 derived dimensions from a sample of $n_M = 4082$ men and $n_F = 1986$ women. Although data were compiled demographically in terms of race, ethnicity, gender, age, and geographic location, the analysis in this paper partitions the data into two samples based exclusively on gender. Of the 93 directly measured attributes and 41 derived attributes acquired from each of the 6068 individuals in the combined sample, the only statistics pertinent to this paper are the weight (converted to mass) and height, from which the sample BMI values are calculated according to Equation (1). Details of the measurement apparatus, measurement procedure, and steps taken to assure accuracy are described in the Technical Report [10].

3.1. Distribution of Height

Figure 5 shows a histogram (gray bars) of the distribution of heights of the male subgroup (left panel) and female subgroup (right panel) in the ANSUR population. Corresponding histograms of the natural logarithm of the heights are shown in Figure 6. Table 2 summarizes the sample statistics obtained from

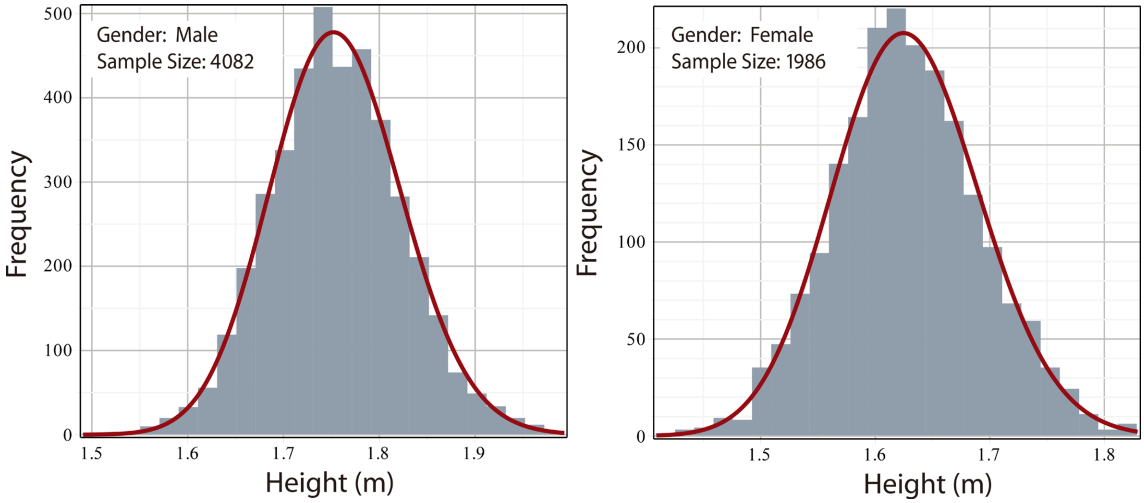


Figure 5. Histograms (gray bars) of the height of male (left) and female (right) soldiers compiled from the ANSUR data. Superposed envelopes (maroon curves) are the exact lognormal probability density functions.

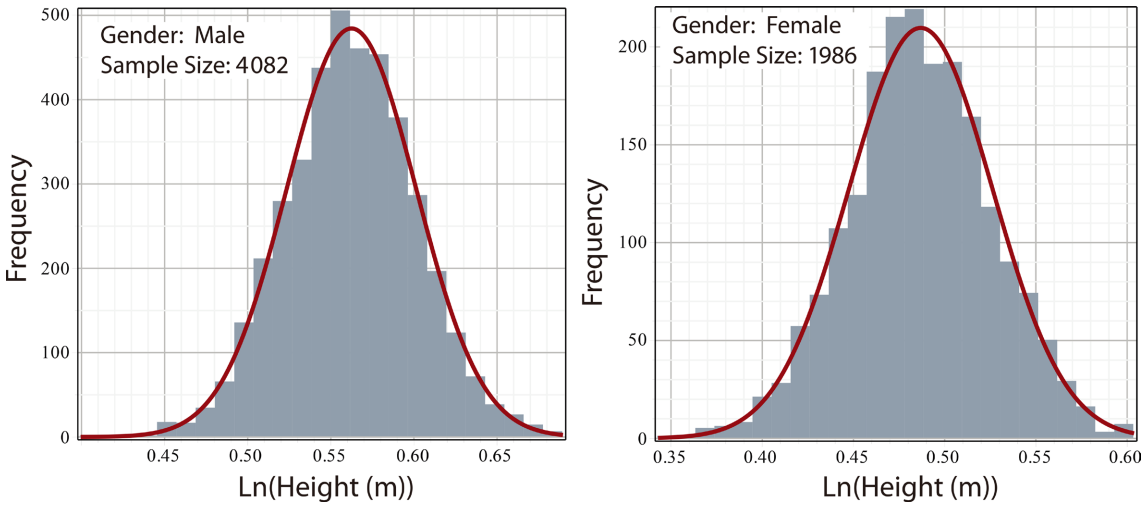


Figure 6. Histograms (gray bars) of the natural logarithm of the height of male (left) and female (right) soldiers derived from the ANSUR data. Superposed envelopes (maroon curves) are the exact Gaussian probability density functions.

analysis of the two sets of data.

The histograms of log-height in **Figure 6** appear symmetric about the mean and can be well fitted by Gaussian profiles with sample means and standard deviations

$$\text{Male Subgroup:} \quad \begin{matrix} m_{HM} = 0.5624 \\ s_{HM} = 0.0390 \end{matrix} \tag{68}$$

$$\text{Female Subgroup:} \quad \begin{matrix} m_{HF} = 0.4869 \\ s_{HF} = 0.0394 \end{matrix} \tag{69}$$

calculated directly from the unpartitioned data (in contrast to partitioning the data into categories and applying a maximum likelihood or Bayesian estimation procedure).

Table 2. Descriptive statistics of sample height, weight, body mass index (BMI).

Statistic	Sample M: 4082 F: 1986	Mean	Standard Deviation	Skewness	Kurtosis	Min Value	Max Value
Height (m)	Male	1.7562	0.0685	0.1113	3.0680	1.49	1.99
	Female	1.6285	0.0642	0.0876	3.0041	1.41	1.83
Weight (kg)	Male	85.5240	14.2190	0.4817	3.3583	39.30	144.20
	Female	67.7582	10.9819	0.5545	3.6599	35.80	119.6
BMI (kg/m^2)	Male	27.6863	4.0390	0.3568	3.1129	15.35	43.45
	Female	25.4960	3.4908	0.5144	3.4797	16.37	40.78
Ln Height	Male	0.5624	0.0390	−0.0090	3.0594		
	Female	0.4869	0.0394	−0.0312	3.0222		
Ln Weight	Male	4.4351	0.1654	−1.4686	3.0295		
	Female	4.2030	0.1604	2.4300	3.1094		
Ln BMI	Male	3.3103	0.1458	−0.0582	2.8854		
	Female	3.2293	0.1354	0.0995	2.9858		
			Standard Error				
Correlation r (lnH & lnW)	Male	0.4716	0.0122				
	Female	0.5387	0.0159				
Correlation ρ (H & W)	Male	0.4689	0.0010				
	Female	0.5335	0.0015				

Chi-square tests of the goodness of fit of the log-height histograms to Gaussian profiles are summarized in Table 3. For $\nu = 24$ degrees of freedom (data partitioned into 25 categories), the tests yielded respective p-values of 35.73% (male) and 58.77% (female). It is to be recalled that the p-value is the probability that a subsequent random sample from the same total population would result in a chi-square value equal to or greater than the observed value, assuming the null hypothesis is correct [35]. The null hypothesis in testing the histograms of Figure 6 is that they are samples from Gaussian distributions with parameters given by Equations (68) and (69). The critical statistic of a chi-square test is the chi-square value beyond which the p-value is below 5%. The p-values in Table 3 are all well above 5%.

The significance of a chi-square test is not that it proves the null hypothesis to be true, but that the null hypothesis cannot be rejected on the basis of the test. Nevertheless, the test supports the inference that, if the histograms of log-height are Gaussian, then the height, itself, is distributed lognormally for both male and female subgroups. This is evidenced in Figure 5 by the superposed lognormal profiles corresponding to the distributions $\Lambda\left(m_{HM}, s_{HM}^2\right)$ for males and $\Lambda\left(m_{HF}, s_{HF}^2\right)$ for females. Chi-square tests of the lognormal fits, reported in Table 3, show

p-values of 41.67% for males and 59.08% for females, which again support the null hypothesis.

Although the visual appearance of the histograms of height in Figure 5, for both male and female subgroups, may suggest that these data are distributed normally, this appearance is deceptive and incorrect, given that the natural logarithm of the set of variates yield normal distributions. By contrast, the natural logarithm of a normal variable is not distributed normally, as shown in Figure 7. The blue profile is a normal (Gaussian) distribution based on the same height parameters $(m=1.8, s=0.5)$ as the example used in Figure 1. The red profile is the distribution of the natural logarithm of the Gaussian variates.

Table 3. Chi-square tests of the log-normal fit to height, weight, and BMI.

Variable	Parameters of PDFs $\Lambda(m,s)$ and $N(m,s)$		Chi-Square Tests with d.o.f $\nu=24$			
			Physical Variable X		Ln of Variable X	
	Location (m)	Scale (s)	$\chi^2_\nu\|_\Lambda$	*P-Value* % Λ	$\chi^2_\nu\|_N$	*P-Value* % N
Height (M)	0.5626	0.0390	24.80	41.67	25.92	35.73
Height (F)	0.4869	0.0394	21.81	59.08	21.86	58.77
Weight (M)	4.4351	0.1654	27.56	27.89	21.36	61.71
Weight (F)	4.2030	0.1604	22.50	54.97	18.92	59.08
BMI (M)	3.3103	0.1458	24.32	44.37	22.98	52.12
BMI (F)	3.2293	0.1354	32.20	12.21	32.02	12.66

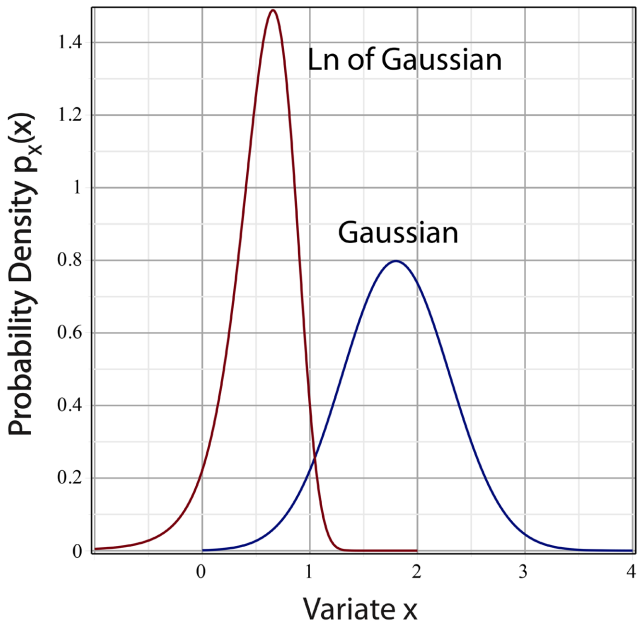


Figure 7. Profile of the PDF of a normal variable $Y=N(1.8,0.05^2)$ (blue) and the profile of the log-of-normal variable (renamed a logGauss variable) $X=\ln(Y)$ (maroon). One sees that a logGauss variable is not distributed normally.

To examine this issue analytically, consider a normal variable $Y = N\left(m, s^2\right)$ and the log-of-normal variable $X = \ln(Y)$. To avoid confusing the term "log-of-normal" with the entrenched designation "lognormal" for a variable whose natural logarithm is normal, we will call X in this example a logGauss random variable. Employing the transformation methods of previous sections, one can readily show that the PDF of a logGauss variable takes the form

$$p_X(x) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{\left(e^x - m\right)^2}{2s^2} + x \right), \tag{70}$$

which is *not* equivalent to the PDF of a normal (Gaussian) distribution. For variates x in the vicinity of the maximum point at $\ln(m)$, one can truncate at first order a Taylor series expansion of the numerator $\left(e^x - m\right)$ in Equation (70) to obtain an approximate PDF of Gaussian form. However, the expansion is not valid at the wings, which descend more quickly than a Gaussian on the right side and extend more slowly and into the nonphysical negative range on the left side.

It is clear, then, that the distribution of heights of males and females in the ANSUR data is not a normal distribution, but, in conformity with our applied statistical tests and the theoretical analyses of [2] [11], is consistent with a lognormal distribution. Moreover, given that the same biological processes are likely to determine height in any population of healthy males or females with access to adequate nutrition, we believe it reasonable to infer that human height in all such populations is distributed lognormally.What distinguishes one population from another would be the parameters, not the form, of the distribution.

3.2. Distribution of Weight (Mass)

Figure 8 shows a histogram (gray bars) of the distribution of weight (converted to mass) of the male subgroup (left panel) and female subgroup (right panel) in the ANSUR population. The mass histograms in **Figure 8** are skewed to the right and are clearly non-Gaussian. Corresponding histograms of the natural logarithm of the masses are shown in **Figure 9**. **Table 2** summarizes the sample statistics obtained from analysis of the two sets of data.

As with the attribute of height in the previous section, the histograms of log-mass in **Figure 9** appear symmetric about the mean and are well fitted by Gaussian profiles with the following sample means and standard deviations

$$\text{Male Subgroup:} \quad \begin{matrix} m_{WM} = 4.4351 \\ s_{WM} = 0.1654 \end{matrix} \tag{71}$$

$$\text{Female Subgroup:} \quad \begin{matrix} m_{WF} = 4.2030 \\ s_{WF} = 0.1604 \end{matrix} \tag{72}$$

calculated directly from the unpartitioned data. (Note: We use the subscript W for weight in relations (71) and (72), even though the distribution function and associated moments are for mass, since weight was the attribute actually measured. Also, we reserve the subscript M to represent "Male".)

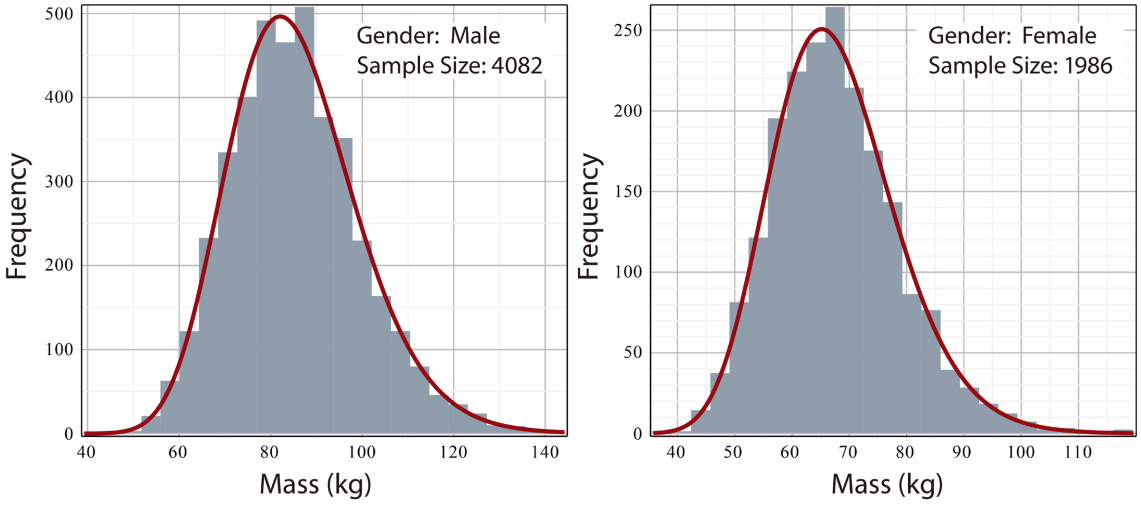


Figure 8. Histograms (gray bars) of the mass of male (left) and female (right) soldiers compiled from the ANSUR data. Superposed envelopes (maroon curves) are the exact lognormal probability density functions.

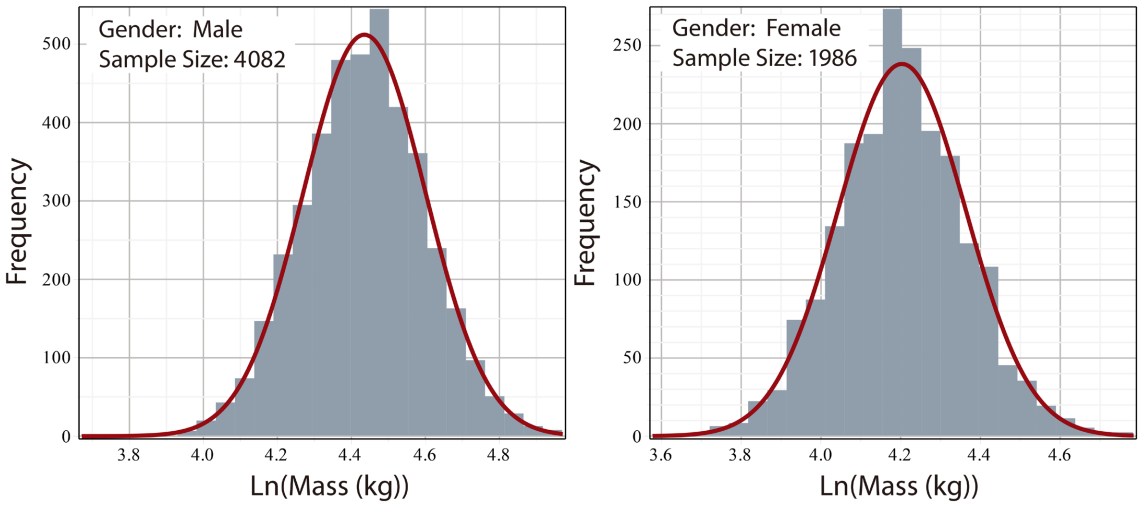


Figure 9. Histograms of the natural logarithm of the mass of male (left) and female (right) individuals derived from the ANSUR data. Superposed envelopes (maroon curves) are the exact Gaussian probability density functions.

Chi-square tests of the goodness of fit of the log-mass histograms in **Figure 9** to Gaussian profiles are summarized in **Table 3**. For $\nu = 24$ degrees of freedom, the tests yielded respective p-values of 61.71% (male) and 59.08% (female). Likewise, chi-square tests of the fit of the mass histograms to lognormal profiles in **Figure 8** yielded p-values of 27.89% (male) and 54.97% (female). Altogether, the chi-square tests of the histograms in **Figure 8** and **Figure 9** well support the null hypothesis that weight (mass) is distributed lognormally in both male and female subgroups of the ANSUR population. As with height, there is reason to infer that the attribute of weight in healthy human populations accessible to adequate nutrition will follow a lognormal distribution.

3.3. Correlation of Height and Weight (Mass)

Figure 10 shows a scatter plot of the weight (converted to mass) against height

for males (left panel) and females (right panel) of the ANSUR sample. Each point in a scatter plot is the mass and height of a single individual. The elongated shapes of the scatter plots clearly demonstrate that the data are linearly correlated. There may also be higher order correlations, but in this paper we are concerned exclusively with linear correlation as quantified by the Pearson correlation coefficients r and ρ defined by Equations (47) and (53), respectively, and predicted by Equation (55) for lognormal distributions.

Figure 11 displays the scatter plots of Figure 10 rescaled by dividing the variates of the two random variables by their sample standard deviations. The resulting variate is a pure number without units or dimensions. Superposed on the

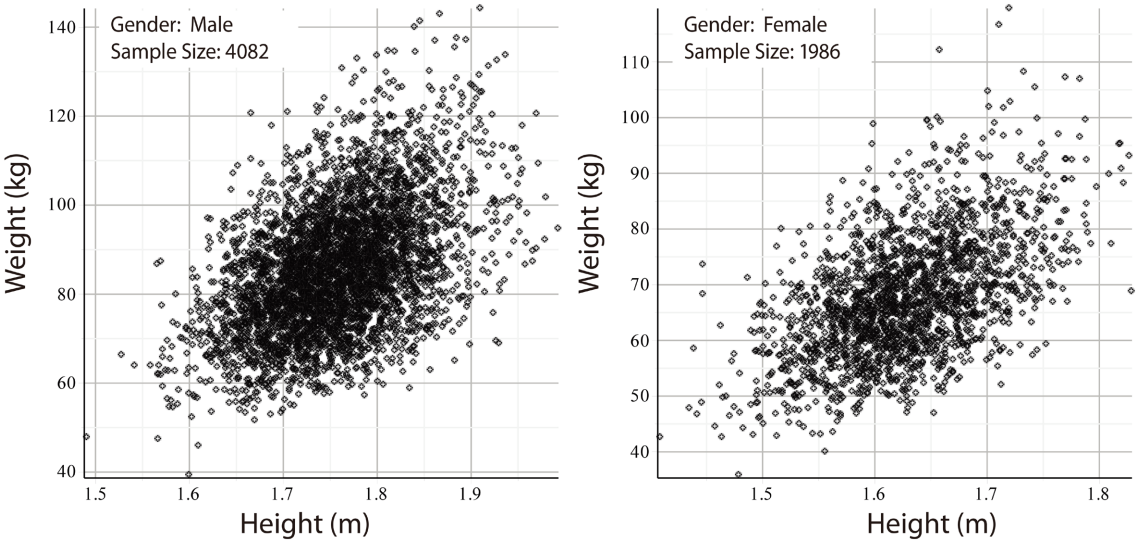


Figure 10. Correlation of mass (kg) and height (m) for males (left panel) and females (right panel) of the ANSUR sample. The elongated scatter patterns display a linear correlation.

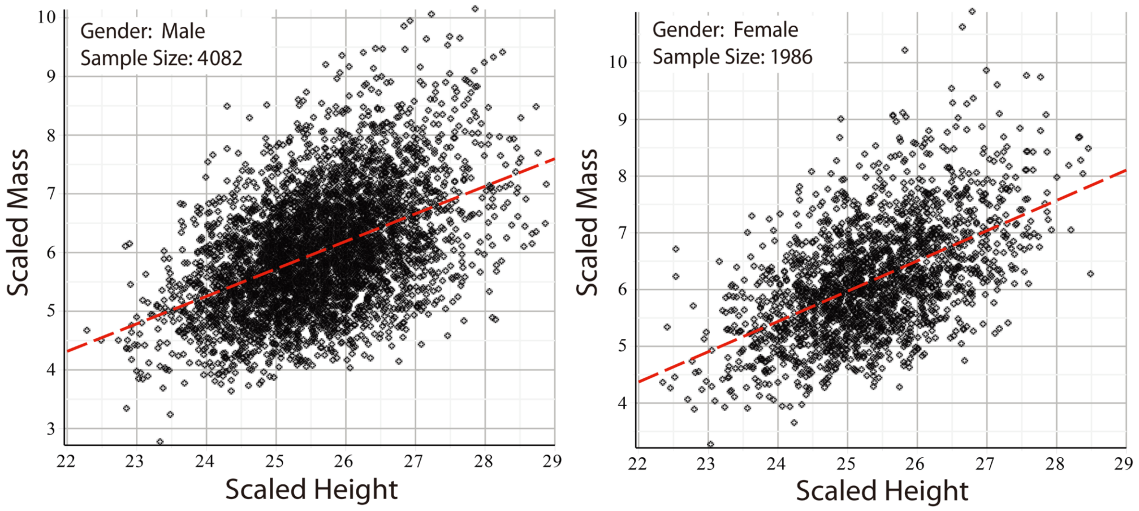


Figure 11. Correlation of mass and height scaled by their respective standard deviations for the data in Figure 10. The scaled variables are pure numbers without units. Each superposed dashedred line is a linear least squares fit to the scaled data. The slope of the left (right) line is precisely the correlation coefficient ρ for males (females) as predicted from lognormal theory (Equation (55)) and shown in Table 2 and Table 4.

dimensionless scatter plots is the line of regression obtained from a least-squares fit to the scaled data. The respective slopes of the lines in the left and right panels accurately yielded the correlation coefficient ρ for males and females, respectively, as recorded in Table 2 and Table 4. For purposes of comparison, Figure 12 shows a simulated scatter plot of *uncorrelated* weight and height, obtained from 10,000 samples drawn independently from lognormal random number generators (RNGs) with the same parameters as given in relations (69) and (72) for the female subgroup in the ANSUR data. The overall shape is circular, apart from fluctuations at the periphery.

It is an important point worth clarifying *why* the slope of the line of regression to the scaled scatter plot is an *exact* geometric representation of the Pearson correlation coefficient. We have not seen this point discussed elsewhere although Galton seems to have understood this point empirically in 1888 [36]. A linear least-squares (LLS) fit with slope a and intercept b

$$y = ax + b \tag{73}$$

to the raw data (*i.e.* the scatter plot of variates y against variates x) leads to the standard LLS slope [32]

$$\hat{a} = \frac{\frac{1}{n}\sum xy - \left(\frac{1}{n}\sum x\right)\left(\frac{1}{n}\sum y\right)}{\frac{1}{n}\sum x^2 - \left(\frac{1}{n}\sum x\right)^2} \tag{74}$$

which is the sample statistic corresponding to the population statistic

$$a = \frac{\langle XY\rangle - \langle X\rangle\langle Y\rangle}{\sigma_X^2}. \tag{75}$$

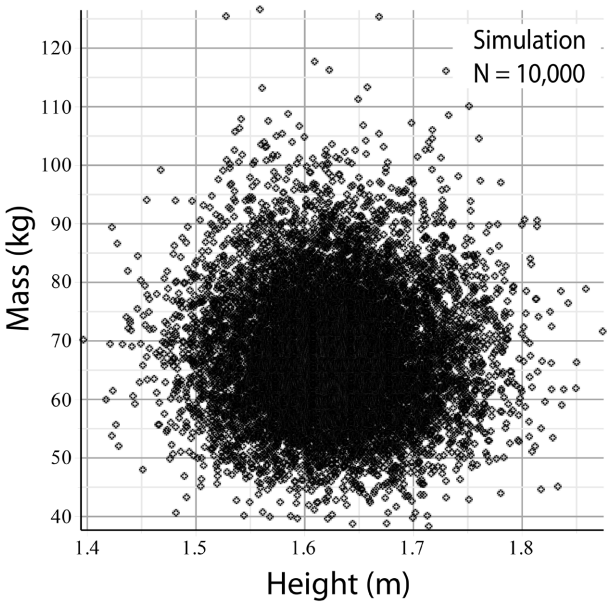


Figure 12. Simulated scatter plot of *uncorrelated* weight and height, obtained from 10,000 samples drawn independently from lognormal random number generators (RNGs) with parameters corresponding to the ANSUR female subgroup.

Relation (75) is *not* the Pearson correlation coefficient expressed by Equation (53). However, if one substitutes into Equation (74) the scaled variables

$$\begin{aligned} x' &\equiv x/\sigma_X \\ y' &\equiv y/\sigma_Y \end{aligned} \tag{76}$$

then it follows straightforwardly that the resulting sample statistic corresponds to the population statistic a'

$$a' = \frac{\langle XY\rangle - \langle X\rangle\langle Y\rangle}{\sigma_X \sigma_Y}, \tag{77}$$

which *is* the Pearson correlation coefficient.

The quantity ρ in Table 4 is especially revealing, for it shows the agreement to three or four decimal places of the values of the empirical weight-height correlation coefficient for male and female subgroups obtained in 3 different ways: 1) direct calculation of the covariance of unpartitioned data displayed in Figure 10; 2) calculation of the slope of the scaled data in Figure 11; and 3) *prediction* of ρ by lognormal theory from the Pearson correlation coefficient r of the bivariate normal distribution $N_{WH}\left(m_W, s_W^2; m_H, s_H^2; r\right)$. Thus, analysis of the Pearson correlation coefficient ρ reinforces the conclusion that weight and height comprise correlated bivariate lognormal random variables symbolized by $\Lambda_{WH}\left(m_W, s_W^2; m_H, s_H^2; r\right)$.

We also note here, in anticipation of the next section, the very close agreement in Table 4 of the sample mean and standard deviation of the log-BMI data with the corresponding values predicted from Equation (64), which again depend on the Pearson correlation r.

3.4. Distribution of Body Mass Index (BMI)

Figure 13 shows histograms of the BMI for males (left panel) and females (right panel) calculated from the weight and height data of the ANSUR sample and normalized to unit area. The dashed red-blue envelope curve in each panel is actually a superposition of two theoretical curves: a) a lognormal profile (red) with mean and variance obtained directly from the unpartitioned set of natural logarithms of the empirical BMI variates; and b) the lognormal profile (blue) from Equation (62) with parameters *predicted* by Equation (64) and Gaussian statistics (68), (69), (71), (72). The perfect superposition of the two theoretical profiles is strong evidence that human weight and height are described by a correlated bivariate lognormal distribution, and that BMI is likewise distributed lognormally with theoretically determined, nonadjustable parameters.

Chi-square tests of the hypothesis that BMI is a lognormal variable is summarized in Table 3. For $\nu = 24$ degrees of freedom, the tests yielded respective p-values of 44.37% for the male subgroup and 12.21% for the female subgroup.

Corresponding histograms of the natural logarithm of BMI are shown in Figure 14, superposed by Gaussian envelope curves computed with the parameters used in Figure 13. Chi-square tests of the goodness of fit, given in Table 3, yielded

Table 4. Comparison of BMI statistics from sampling and Log-Normal Theory.

Statistic of BMI Distribution	Sample Male (4082)	LN Theory (M) $\Lambda(m_W, m_H, s_W, s_H, r)$	Sample Female (1986)	LN Theory (F) $\Lambda(m_W, m_H, s_W, s_H, r)$
Log-Normal Parameters				
Mean m_W	4.4351		4.2031	
Mean m_H	0.5624		0.4869	
SD s_W	0.1654		0.1604	
SD s_H	0.0390		0.0394	
Correlation of lnW & lnH r_{WH}	0.4716		0.5387	
Correlation of W & H ρ_{WH}	0.4689	0.4689	0.5335	0.5359
BMI Statistics				
Mean $\ln(B)$ m_B	3.3103	3.3099	3.2293	3.2292
SD $\ln(B)$ s_B	0.1458	0.1432	0.1354	0.1354
Mean	27.6863	27.6873	25.4960	25.4948
SE Mean	0.06322	0.0635	0.0783	0.0778
Variance	16.3133	16.4762	12.1856	12.0216
SE Variance	0.3711	0.3954	0.4306	0.4092
Skewness	0.3568	0.4430	0.5144	0.4105
SE Skewness	0.0643	0.0745	0.1112	0.1040
Kurtosis	3.1129	3.3509	3.4797	3.3011
SE Kurtosis	0.1724	0.2512	0.3670	0.3404

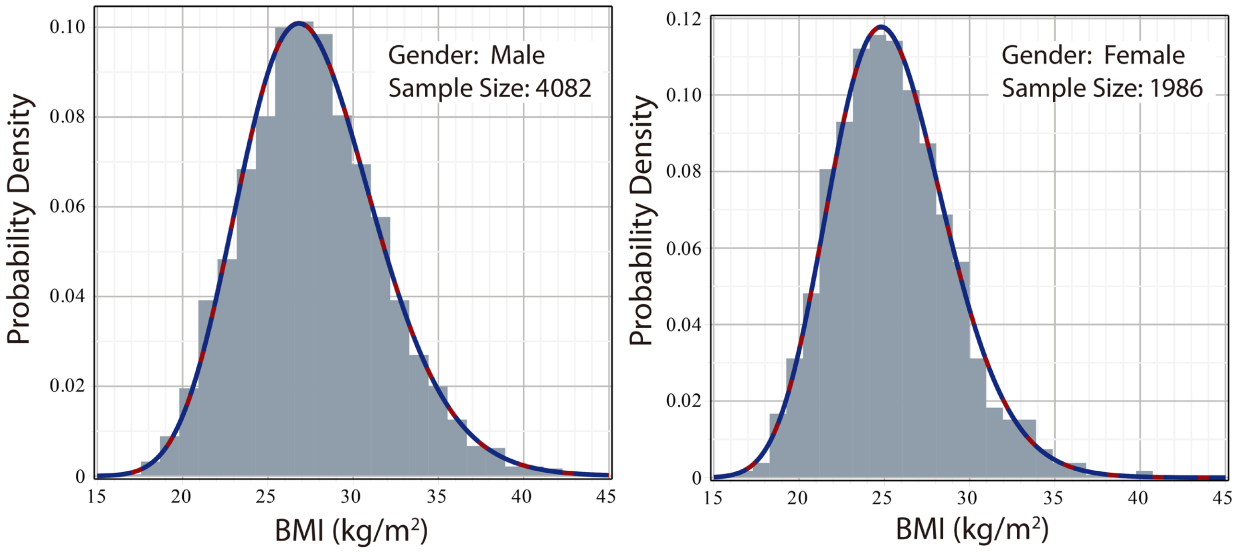


Figure 13. Histograms (gray bars) of the BMI for males (left panel) and females (right panel) calculated from the weight and height data of the ANSUR sample and normalized to unit area. The dashed red-blue envelope in each panel is a superposition of two probability density profiles: (a) a lognormal profile (red) with parameters (mean and variance) obtained directly from the unpartitioned natural logarithms of the empirical BMI variates; and (b) the lognormal profile (blue) from Equation (62) with parameters *predicted* from Equation (64).

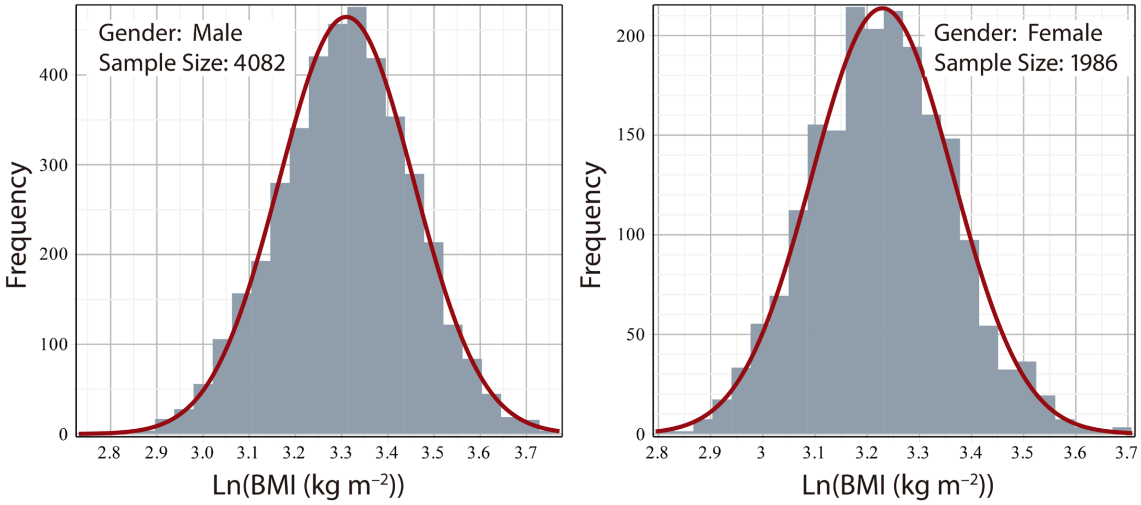


Figure 14. Histogram (gray bars) of the natural logarithm of the BMI of male (left) and female (right) individuals in the ANSUR sample. The superposed maroon profile in each panel is the theoretical normal PDF with Gaussian parameters predicted by lognormal theory, Equation (64).

p-values of 52.12% for the male subgroup and 12.66% for the female subgroup.

It is to be emphasized that the excellent fit of the theoretical probability density to the normalized histograms of BMI depends crucially on the correlation of the two variables, weight and height. Recall that the 4 parameters (2 pairs of means and variances) that separately characterize the lognormal distributions of weight (mass) and height are obtained experimentally from the *marginal* distributions of what is actually a bivariate normal distribution $N_{WH}\left(m_W, s_W^2; m_H, s_H^2; r\right)$. However, the marginal distributions are independent of the Pearson correlation parameter r. If one is ignorant of, or intentionally disregards, the correlation of weight and height, the resulting theoretical probability density function may then fit the observed BMI distribution *very poorly*, as illustrated in **Figure 15**.

The gray translucent bars in **Figure 15** comprise the BMI histogram of the ANSUR female subgroup displayed in **Figure 13**. The black enveloping curve is the theoretical PDF, Equation (62), with empirical correlation $r = 0.5387$, also shown in **Figure 13**. By contrast, the orange bars comprise a normalized histogram of BMI simulated by 10,000 samples drawn independently from RNGs for mass and height. The magenta envelope is the theoretical PDF, Equation (62), with $r = 0$. The histogram composed of uncorrelated samples of weight and height is much wider that the true (*i.e.* empirically obtained) histogram, and of lower maximum (since the total area under a normalized histogram is unity). As shown in **Figure 15**, the tails of the two histograms, which characterize the subpopulations at greatest risk of obesity and metabolic disease, differ significantly. To disregard positive (negative) correlation of weight and height is to significantly overcount (undercount) the population at greatest risk.

4. Conclusions

The body mass index (BMI) is one of the most widely employed medical risk

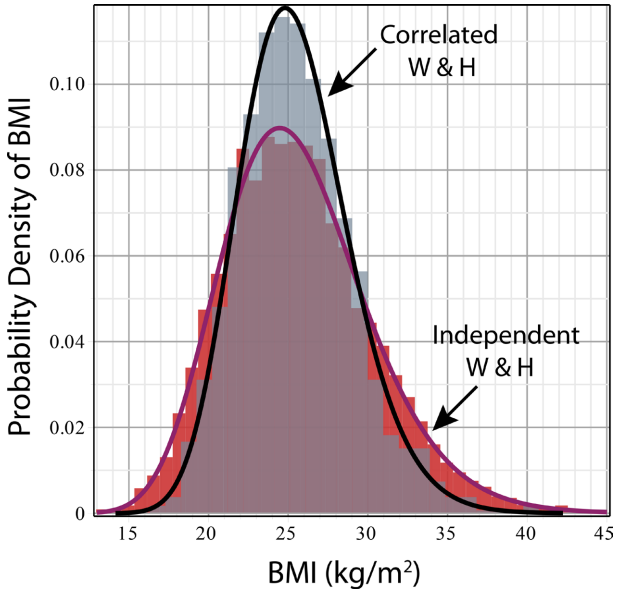


Figure 15. Comparison of the histogram of BMI in the ANSUR female subgroup (translucent gray bars; sample size = 1986) with a simulated histogram (orange bars, sample size = 10,000) obtained from lognormal random number generators (RNGs) programmed with the same means and variances for weight (mass) and height as in the ANSUR sample. Weight and height variates are correlated in the ANSUR sample, but are drawn from independent RNGs in the simulation. Superposed on the two normalized histograms are the theoretical PDF Equation (62) with Pearson $r = 0.5387$ for the ANSUR sample (black curve) and $r = 0$ for the simulation (magenta curve).

factors in current use, given the epidemic proportions of obesity among populations of both industrialized and developing countries. A significant amount of research over many years has been devoted to modeling and/or approximating an empirical distribution function for BMI. In this paper, we derived by rigorous statistical reasoning the mathematically *exact* form of the probability density function (PDF), Equation (61) to which the definition of BMI as the ratio of mass to the square of height inexorably leads. This PDF is uniquely determined by the correlated bivariate distribution of weight and height, the form of which we deduced from a large anthropometric data base.

The advantage of an exact theory over an empirically matched mathematical expression is that the exact theory is valid over the entire allowed range of its variables and applies to other statistical populations than the one (or few) used for purposes of testing and confirmation. By contrast, an expression obtained by curve-fitting has a limited range of validity and cannot be relied on to characterize other statistical populations. Perhaps even more significant is that the exact theory provides insights into the relationships of its variables, whereas an approximate expression found by curve fitting merely provides at best a numerical or graphical coincidence without an underlying scientific basis.

We proposed theoretically and demonstrated experimentally by statistical analysis of a large anthropometric data base that human weight and height constitute a correlated bivariate lognormal distribution represented by

$\Lambda_{WH}\left(m_W, s_W^2; m_H, s_H^2; r\right)$. The five parameters defining the PDF (2 means, 2 variances, and 1 linear correlation), inferred from the natural logarithm of the mass and height variates, uniquely predict the BMI PDF (62) from which all statistical moments of BMI follow. There are no freely adjustable parameters in the exact PDF. From the resulting form of the exact BMI PDF, we established that BMI is rigorously a lognormal random variable itself.

Our investigation of the correlation of weight and height has shown that it can strongly affect the BMI PDF and statistical moments, particularly in regard to the amplitude and extent of the tail of the distribution, which relates to the subgroup of a population at greatest risk. In particular, a positive (negative) linear correlation leads to a narrower (wider) BMI distribution and lower (higher) proportion of high-risk individuals compared with the distribution based on statistically independent weight and height.

In summary, we conclude that a correct and accurate theoretical analysis of the distribution of BMI must include not only the means and variances obtained from the marginal distributions of weight and height, but also a correlation analysis of the two sets of variates. With a complete set of the 5 parameters that define the bivariate weight-height distribution for *each specified demographic*, one would then be in a position to make valid inferences regarding population-specific BMI quantiles (or other statistical measures) that affect public health policy and clinical treatment of individuals.

Acknowledgments

One of the authors (MPS) thanks Trinity College for partial support through the research fund associated with the George A. Jarvis Chair of Physics.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

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

[2] Silverman, M.P. (2019) Crowdsourced Sampling of a Composite Random Variable: Analysis, Simulation, and Experimental Test. *Open Journal of Statistics*, **9**, 494-529.
https://doi.org/10.4236/ojs.2019.94034

[3] WHO (2022) Body Mass Index.
https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi

[4] WHO (2021) Obesity and Overweight.
https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[5] 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

[6] Millar, W.J. (1986) Distribution of Body Weight and Height: Comparison of Esti-

mates Based on Self-Reported and Observed Measures. *Journal of Epidemiology and Community Health*, **40**, 319-323. https://doi.org/10.1136/jech.40.4.319

[7] Penman, A.D. and Johnson, W.D. (2006) The Changing Shape of the Body Mass Index Distribution Curve in the Population: Implications for Public Health Policy to Reduce the Prevalence of Adult Obesity. *Preventing Chronic Disease A*, **3**, 74. https://www.cdc.gov/pcd/issues/2006/jul/pdf/05_0232.pdf

[8] Ng, M., Liu, P., Thomson, B. and Murray, C.J.L. (2016) A Novel Method for Estimating distributions of Body Mass Index. *Population Health Metrics*, **14**, 1-7. https://doi.org/10.1186/s12963-016-0076-2

[9] Yu, K., Xi, L., Alhamzawi, R., Becker, F. and Lord, J. (2018) Statistical Methods for Body Mass Index: A Selective Review. *Statistical Methods in Medical Research*, **27**, 798-811. https://doi.org/10.1177/0962280216643117

[10] 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

[11] Silverman, M.P. (2019) Extraction of Information from Crowdsourcing: Experimental Test Employing Bayesian, Maximum Likelihood, and Maximum Entropy Methods. *Open Journal of Statistics*, **9**, 571-600. https://doi.org/10.4236/ojs.2019.95038

[12] 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

[13] Silverman, M.P. (2014) A Certain Uncertainty: Nature's Random Ways. Cambridge University Press, Cambridge, 17-18, 28-32, 54-61, 272-327. https://doi.org/10.1017/CBO9781139507370.006

[14] Mood, A.M., Graybill, F.A. and Boes, D.C. (1974) Introduction to the Theory of Statistics. 3rd Edition, McGraw-Hill, New York, 181-188, 198-212.

[15] Hald, A. (1952) Statistical Theory with Engineering Applications. Wiley, New York, 159-174.

[16] Jaynes, E.T. (1957) Information Theory and Statistical Mechanics. *Physical Review*, **106**, 620-630. https://doi.org/10.1103/PhysRev.106.620

[17] Cypress, A.M. (2022) Reassessing Human Adipose Tissue. *NEJM*, **386**, 768-779. https://doi.org/10.1056/NEJMra2032804

[18] CDC (2021) About Adult BMI. https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

NHLBI Obesity Education Initiative Expert Panel (1998) Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults. NIH Publication No. 98-4083.

[19] Weir, C.B. and Arif, J. (2021) BMI Classification Percentile and Cut off Points. StatPearls Publishing, Treasure Island, 1-5. https://www.ncbi.nlm.nih.gov/books/NBK541070

[20] WHO Expert Consultation (2004) Appropriate Body-Mass Index for Asian Populations and Its Implications for Policy and Intervention Strategies. *The Lancet*, **363**, 157-163. https://doi.org/10.1016/S0140-6736(03)15268-3

[21] Fangjian, G. and Garvey, W.T. (2016) Cardiometabolic Disease Risk in Metabolically Healthy and Unhealthy Obesity: Stability of Metabolic Health Status in Adults. *Obesity*, **24**, 516-525. https://doi.org/10.1002/oby.21344

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

[23] The Editors (2020) Weight Is Not Enough. *Scientific American*, **322**, 10.

[24] Rose, G. (1981) Strategy of Prevention: Lessons from Cardiovascular Disease. *British Medical Journal*, **282**, 1847-1851. https://doi.org/10.1136/bmj.282.6279.1847

[25] Rose, G. (1992) The Strategy of Preventive Medicine. Oxford University Press, New York.

[26] Hoffman, A. and Vandenbroucke, J.P. (1992) Geoffrey Rose's Big Idea, *BMJ*, **305**, 1519-1520. https://doi.org/10.1136/bmj.305.6868.1519

[27] Arfken, G.B. and Weber, H.J. (2005) Mathematical Methods for Physicists. 6th Edition, Elsevier, New York, 83-85, 669-670, 975.

[28] Chou, Y. (1969) Statistical Analysis with Business and Economic Applications. Holt, Rinehart, and Winston, New York, 218-222.

[29] Kendall, M.G. and Stuart, A. (1963) The Advanced Theory of Statistics Vol. 1 Distribution Theory. Hafner, New York, 333-334.

[30] Gumbel, E.J. (1958) Statistics of Extremes. Echo Point Books & Media, Brattleboro, 1-6. https://doi.org/10.7312/gumb92958

[31] Hogg, R.V., McKean, J.W. and Craig, A.T. (2005) Introduction to Mathematical Statistics. Prentice Hall, Upper Saddle River, 101-106, 174-175.

[32] Hoel, P.G. (1947) Introduction to Mathematical Statistics. Chapman & Hall, London, 78-84.

[33] 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

[34] Taylor, J.R. (1997) An Introduction to Error Analysis. 2nd Edition, University Science Books, Sausalito, 146-147.

[35] Altman, D.G. (1999) Practical Statistics for Medical Research. Chapman & Hall/CRC, New York, 167-171.

[36] Stigler, S.M. (1989) Francis Galton's Account of the Invention of Correlation. *Statistical Science*, **4**, 73-79. https://doi.org/10.1214/ss/1177012580

Appendix

CDC—Centers for Disease Control and Prevention of the U.S. National Institutes of Health
BMJ—British Medical Journal
JAMA—Journal of the American Medical Association
NCBI—National Center for Biotechnology Information
NEJM—New England Journal of Medicine
NHANES—National Health and Nutrition Examination Survey
NHLBI—National Heart Lung and Blood Institute
NIH—National Institutes of Health
NLM—National Library of Medicine
WHO—World Health Organization