STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:3109–3125

Sympercents: symmetric percentage differences on the 100 $\log_e$ scale simplify the presentation of log transformed data

T. J. Cole*,†

Department of Paediatric Epidemiology and Biostatistics, Institute of Child Health, London WC1N IEH, U.K.

SUMMARY

The results of analyses on log transformed data are usually back-transformed and interpreted on the original scale. Yet if natural logs are used this is not necessary – the log scale can be interpreted as it stands. A difference of natural logs corresponds to a fractional difference on the original scale. The agreement is exact if the fractional difference is based on the logarithmic mean. The transform $y = 100 \log_e x$ leads to differences, standard deviations and regression coefficients of y that are equivalent to symmetric percentage differences, standard deviations and regression coefficients of x. Several simple clinical examples show that the 100 $\log_e$ scale is the natural scale on which to express percentage differences. The term sympercent or s% is proposed for them. Sympercents should improve the presentation of log transformed data and lead to a wider understanding of the natural log transformation. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

Despite their widespread use in statistics, logarithms can be a mystery to non-specialists [1, 2]. The justification for a logarithmic transformation is unfamiliar, be it non-linearity, heteroscedasticity or non-Normality, but the unintuitive nature of the log *scale* is probably the biggest difficulty.

Statisticians of course think differently. They are comfortable with the log transformation as a statistical tool, and they also know that it is special [3] – the only power transformation where a difference on the transformed scale is interpretable on the original scale as a ratio [4, 5].

To simplify the presentation of data analysed on the log scale, the standard procedure is to antilog the results [4, 6, 7]. This ensures that the base of logarithms used, common or natural logs, is irrelevant. Standard statistics texts often imply that natural logs have advantages over common logs [4, 6, 7], but they do not spell out what the advantages are and use examples with logs to both bases.

This ambivalence about the relative merits of natural and common logs indicates a widespread ignorance about the benefits of the natural log scale. This paper aims to explain, to the specialist and non-specialist alike, why the natural log scale is 'natural' as well as 'special'. Though

* Correspondence to: T. J. Cole, Department of Paediatric Epidemiology and Biostatistics, Institute of Child Health, 30 Guilford Street, London WC1N 1EH, U.K.
† E-mail: Tim.Cole@ich.ucl.ac.uk.

Received October 1998
Accepted November 1999

developed independently, the work here builds on ideas from the econometric literature of index numbers, well summarized by the important and sadly under-cited paper of Törnqvist *et al.* [8]. The concept does not seem to have been described in the medical statistics literature.

The natural log scale is a scale on which differences can be interpreted directly, without back-transformation to the original scale. Natural log differences correspond to fractional differences on the original scale. Multiplied by 100, natural log differences are equivalent to symmetric percentage differences, that is, percentage differences calculated with the mean of the two numbers as the denominator. This equivalence extends to other summary statistics on the 100 natural log scale, such as log standard deviations, equivalent to coefficients of variation, log regression coefficients, related to fractional regression coefficients, and logs of rate ratios expressed as symmetric percentage differences.

The paper is structured as follows: Section 2 discusses logarithms and describes a symmetric and additive form of the percentage difference. Section 3 develops the algebra relating logarithmic and fractional differences, standard deviations and regression coefficients by introducing the logarithmic mean. Section 4 has a series of simple practical examples, and Section 5 provides the discussion and conclusions. Non-specialists are encouraged to skip Section 3.

2. BACKGROUND

2.1. Logarithms

The $\log_{10}$ or common log transformation is useful for expressing large and small numbers compactly, for example, the common logs of one million and one millionth are 6 and -6, respectively. The integer part of the log gives the approximate size of the original number as a power of 10, and to within an order of magnitude common logs can be antilogged by inspection. By contrast, the fractional part of the common log corresponds to a number between 1 and 10, which cannot be interpreted by inspection. The same applies to the difference between two numbers on the common log scale – their ratio cannot easily be inferred from the log difference, unless of course they differ by some orders of magnitude.

Natural logs are logs to base e, and are denoted by $\log_e$ or ln. They are the opposite of common logs in that on their own they are uninterpretable, but as differences they are simple to interpret. The difference between the natural logs of two numbers is the fractional difference between the numbers.

This distinction between common and natural logs is relevant for the choice of base when plotting log transformed data. If the data values themselves are of interest, and extend across several orders of magnitude, then logs to base 10 are preferable. Obvious examples are inter-country comparisons of gross national product or bacterial concentrations in microbiology. Conversely, if differences between the data are more relevant then natural logs are better, and multiplied by 100 they can be viewed as a percentage scale. As an example, logged serial data can be plotted with the initial point at the origin, and this presents the data in a form analogous to per cent of the baseline.

A word about notation. From here on, all logarithms are calculated to base e. Phrases such as 'log difference' or 'log SD' are to be read as meaning the difference or SD *calculated on the log scale*. This usage is consistent with the term 'log Normal distribution', meaning a distribution that is Normal on the log scale.

2.2. Percentage differences, symmetry and additivity

Conventionally the percentage difference between two numbers is the percentage change from x_1 to x_2, defined as

$$\text{Percentage change from } x_1 \text{ to } x_2 = 100\,\frac{x_2 - x_1}{x_1}$$

If the order is reversed, the percentage change from x_2 to x_1 is $100\,[(x_1 - x_2)/x_2]$. The numerator changes sign, and the denominator changes from x_1 to x_2. Thus unless x_1 and x_2 are equal, the two percentages differ in absolute magnitude. Compare this with the difference between x_1 and x_2, which is unchanged apart from the sign if x_1 and x_2 are switched: $(x_2 - x_1)$ or $(x_1 - x_2)$. The percentage difference is not *symmetric* – swapping x_1 and x_2 changes not only its sign but also its magnitude.

Take for example the mean heights of British adults aged 20 years: 177.3 cm for men and 163.6 cm for women [9]. Women are $100 \times [(177.3 - 163.6)/177.3] = 100 \times (13.7/177.3) = 7.7$ per cent shorter than men, while men are $100 \times (13.7/163.6) = 8.4$ per cent taller than women. The two percentages are not the same. This is a fundamental problem with the definition of percentage differences. If there is an obvious ordering of x_1 and x_2, for example, time, then the conventional percentage change calculation is reasonable, but if not, as here, the calculation is flawed. Why should the result depend on which way round it is calculated?

More importantly, how can the percentage difference be made symmetric? The simplest way is to average the two percentage changes, giving a percentage difference that is unchanged in absolute value if the two numbers are exchanged. Algebraically

$$\begin{aligned}\text{Mean percentage difference} &= \frac{1}{2}\left(100\,\frac{x_2 - x_1}{x_1} + 100\,\frac{x_2 - x_1}{x_2}\right)\\ &= 100(x_2 - x_1)\,\frac{1}{2}\left(\frac{1}{x_1} + \frac{1}{x_2}\right)\\ &= 100\,\frac{(x_2 - x_1)}{\mathrm{HM}(x)} \qquad (1)\end{aligned}$$

where $\mathrm{HM}(x)$ is the harmonic mean of x_1 and x_2. In the example the harmonic mean height is 170.17 cm, and the mean percentage difference between the sexes is $100 \times (13.7/170.17) = 8.05$ per cent.

Thus the arithmetic mean of the percentages is related to the harmonic mean of the two numbers in (1). The converse also holds – the harmonic mean of the percentages involves the arithmetic mean $\mathrm{AM}(x)$ of the numbers. The arithmetic mean height is 170.45 cm and the harmonic mean difference is 8.04 per cent. Another alternative is the geometric mean of the two percentages, which gives the geometric mean $\mathrm{GM}(x)$ in (1), and which at 170.31 cm gives a difference of 8.044 per cent. Thus alternative forms of the mean percentage in (1) lead to different means in the denominator.

Apart from asymmetry the conventional percentage difference has another unsatisfactory feature – it is not *additive*. The sum of two or more percentage differences differs from the combined percentage difference. For example, three successive increases of 8 per cent represent an increase overall of 26 per cent not 24 per cent, so 8 per cent $\times 3 \neq 24$ per cent. We know *why* this is so, but it does not occur on the original scale, where $8 \times 3 = 24$, so why should we accept it on the percentage scale?

Percentages ought to be both symmetric and additive, in the same way that absolute quantities are. Logarithms are useful here, as they convert a ratio to a difference. A fundamental property of logarithms to any base is that a difference on the log scale is equal to the log of the corresponding ratio on the original scale: $\log(x_2) - \log(x_1) = \log(x_2/x_1)$. It can be shown that this leads to the required properties of symmetry and additivity. In addition, natural logarithms have the added advantage of giving results in units that are directly interpretable as fractions.

Put briefly, the percentage difference in (1) can be calculated symmetrically and additively as $(100 \log_e x_2 - 100 \log_e x_1)$. The next section justifies this statement.

3. METHODS

3.1. *Fractional differences*

Generalizing from (1), consider an unknown fractional function $f(x)$ such that the symmetric fractional difference between x_1 and x_2 is given by:

$$f(x_2) - f(x_1) = \frac{x_2 - x_1}{\mathrm{ML}(x)} \tag{2}$$

where $\mathrm{ML}(x)$ is a general measure of location based on x_1 and x_2. $\mathrm{ML}(x)$ includes the conventional harmonic, arithmetic and geometric means – $\mathrm{HM}(x)$, $\mathrm{AM}(x)$ and $\mathrm{GM}(x)$ – but is not restricted to them.

Assume that $x_1 = (x - \delta x/2)$ and $x_2 = (x + \delta x/2)$, where $\delta x = (x_2 - x_1)$ and is small. Substituting into (2) gives the fractional difference $\delta f = f(x_2) - f(x_1) = \delta x/x$ since $\mathrm{ML}(x) \rightarrow x$ as $\delta x \rightarrow 0$. In the limit $\mathrm{d}f/\mathrm{d}x = 1/x$ and $f = \ln x$. Thus if $x_1 \approx x_2$

$$\ln x_2 - \ln x_1 \approx \frac{x_2 - x_1}{\mathrm{ML}(x)}$$

The symmetric fractional difference is the same as the difference of the logs of the original numbers. To what extent does this hold in the more general case, when δx is not small? And how critical is the choice of the measure of location?

The function

$$(\ln x_2 - \ln x_1) - \frac{x_2 - x_1}{\mathrm{ML(x)}} \tag{3}$$

is the discrepancy between the log difference and the fractional difference in (2). It vanishes when $\mathrm{ML}(x)$ is equal to the logarithmic mean $\mathrm{LM}(x)$ [10], which is defined as

$$\begin{aligned} \mathrm{LM}(x) &= \frac{x_2 - x_1}{\ln x_2 - \ln x_1} \qquad && x_2 \neq x_1 \\ &= x_1 && x_2 = x_1 \end{aligned} \tag{4}$$

With this measure of location, the fractional difference and the log difference are the same by definition, whatever the size of δx.

The logarithmic mean is an unfamiliar construct, first described in 1972 by Carlson, who showed that $\mathrm{GM}(x) \leqslant \mathrm{LM}(x) \leqslant \mathrm{AM}(x)$ [10]. It is instructive to compare it to other measures of location in more detail. Multiplying (3) by $\mathrm{ML}(x)/x_1$ and substituting $\delta = (x_2 - x_1)/x_1$ (the conventional fractional change from x_1 to x_2) leads to the function

$$\ln(1+\delta)\left[\frac{\mathrm{ML}(x)}{x_1}\right] - \delta \qquad (5)$$

which given $\mathrm{ML}(x)$ can be expanded as a Taylor series in δ.

The three conventional means are introduced by defining $\mathrm{ML}(x)$ as their weighted mean:

$$\begin{aligned}\mathrm{ML}(x) &= a\mathrm{AM}(x) + g\mathrm{GM}(x) + h\mathrm{HM}(x)\\ &= a\frac{x_1+x_2}{2} + g\sqrt{(x_1x_2)} + h\frac{2x_1x_2}{x_1+x_2}\end{aligned} \qquad (6)$$

where $(a+g+h) = 1$.

Substituting (6) into (5), expanding the ln and power terms in powers of δ to the fifth term, and collecting the power coefficients, leads to

$$-\left(\frac{3g+6h-2}{24}\right)(\delta^3 - \delta^4) - \left(\frac{215g+400h-144}{1920}\right)\delta^5 + \mathrm{O}(\delta^6) \qquad (7)$$

Different forms of $\mathrm{ML}(x)$ are compared by substituting the relevant values of g and h in (7). For example, the geometric mean $\mathrm{GM}(x)$ corresponds to $[g=1;\ h=0]$, and reduces (7) to $-\delta^3/24 + \mathrm{O}(\delta^4)$. Table 1 gives the value of $(\mathrm{ML}(x) - \mathrm{LM}(x))$ for three different cases where x_2 is respectively 10 per cent, 100 per cent and 1000 per cent greater than x_1 under the conventional definition.

Table I shows that of the three conventional means $\mathrm{GM}(x)$ is the closest to $\mathrm{LM}(x)$, followed by $\mathrm{AM}(x)$ then $\mathrm{HM}(x)$. The discrepancies are small; for $x_1 = 1$ and $x_2 = 1.1$ (10 per cent difference) $\mathrm{HM}(x)$ differs from $\mathrm{LM}(x)$ by only -0.0016, while for $x_1 = 1$ and $x_2 = 2$ (100 per cent difference) $\mathrm{GM}(x)$ is only 0.028 less than $\mathrm{LM}(x)$. Thus for modest percentage differences, the logarithmic mean and the other means are essentially the same.

Even so there are better forms of $\mathrm{ML}(x)$. The δ^3 and δ^4 terms in (7) vanish when $(g+2h) = 2/3$, and the δ^5 term vanishes for $[g = 182/270;\ h = -1/270]$. The simple set $[g = 2/3;\ h = 0]$ provides a remarkably good fit to (7), corresponding to the weighted mean

$$\mathrm{WM}(x) = \frac{2\mathrm{GM}(x) + \mathrm{AM}(x)}{3} \qquad (8)$$

which is up to three orders of magnitude better than $\mathrm{GM}(x)$ (Table I). Carlson [10] described the same mean.

Generalizing further, $\mathrm{AM}(x)$, $\mathrm{HM}(x)$ and $\mathrm{GM}(x)$ are members of the family of *algebraic* means defined by

$$\mathrm{BM}(x|p) = \left[\frac{1}{n}\sum_{i=1}^{n} x^p\right]^{1/p} \qquad (9)$$

for sample size n, where $\mathrm{BM}(x|p)$ is the back-transformed arithmetic mean of x^p. Thus $\mathrm{AM}(x)$, $\mathrm{GM}(x)$ and $\mathrm{HM}(x)$ correspond to $\mathrm{BM}(x|1)$, $\mathrm{BM}(x|0)$ and $\mathrm{BM}(x|-1)$, respectively, treating the

Table I. Comparing the logarithmic mean LM with other measures of location ML in (5) and (10), where $\delta = (x_2 - x_1)/x_1$.

Measure of location	Formula	Leading δ term in	(ML $-$ LM) when $x_1 = 1$		
			$x_2 = 1.1$	$x_2 = 2$	$x_2 = 11$
(ML)		(5) or (10)	ML $= 1.0492$	ML $= 1.443$	ML $= 4.17$
Harmonic (H(M)	$\dfrac{2x_1x_2}{x_1 + x_2}$	$\dfrac{-\delta^3}{6}$	-1.6×10^{-3}	-1.1×10^{-1}	-2.3
Arithmetic (AM)	$\dfrac{x_1 + x_2}{2}$	$\dfrac{\delta^3}{12}$	7.9×10^{-4}	5.7×10^{-2}	1.8
Geometric (GM)	$\sqrt{(x_1x_2)}$	$\dfrac{-\delta^3}{24}$	-4.0×10^{-4}	-2.8×10^{-2}	-8.5×10^{-1}
Weighted (WM)	$\dfrac{2\,\mathrm{GM} + \mathrm{AM}}{3}$	$\dfrac{\delta^5}{2880}$	3.0×10^{-8}	-1.1×10^{-4}	4.1×10^{-2}
Cube root (CRM)	$\left[\dfrac{\sqrt[3]{x_1} + \sqrt[3]{x_2}}{2}\right]^3$	$\dfrac{\delta^5}{6840}$	1.3×10^{-8}	5.1×10^{-5}	-1.8×10^{-2}
Logarithmic (LM)	$\dfrac{x_2 - x_1}{\ln x_2 - \ln x_1}$	—	0	0	0

log transform as power 0 [11]. $\mathrm{BM}(x|p)$ is an unbiased estimate of the median of x when x^p is Normally distributed.

To see if $\mathrm{BM}(x|p)$ fits (3) better than the weighted mean (8) for some p, substitute (9) with $n = 2$ into (5) to give

$$\ln(1 + \delta)\left[\frac{1 + (1 + \delta)^p}{2}\right]^{1/p} - \delta \qquad (10)$$

In the light of (8) an obvious p-value is 1/3, the mean of 0 for $\mathrm{GM}(x)$ and 1 for $\mathrm{AM}(x)$ weighted in the ratio 2:1. This also simplifies the series expansion of (10) since $1/p = 3$. Again the terms in δ^2, δ^3 and δ^4 vanish, and the term in δ^5 is smaller than for (8) (Table I). Thus the algebraic mean based on the power 1/3 (call it the cube root mean, $\mathrm{CRM}(x)$) is an even better approximation to $\mathrm{LM}(x)$ than the weighted mean. Diewert [12] also described the cube root mean.

The optimal value for p in (10) for each of the examples in Table I, found by direct search, is slightly less than 1/3: 0.333322 for $x_2 = 1.1$, 0.3327 for $x_2 = 2$ and 0.327 for $x_2 = 11$, but 1/3 is clearly close to optimal. Thus the cube root mean is virtually identical to the logarithmic mean.

The practical conclusion is that the means are all essentially the same, so that the log difference is a form of fractional difference. In detail, the mean calculated on the cube root scale is closest to the logarithmic mean, which suggests some deeper significance of the cube root transformation.

3.2. Fractional standard deviations

The previous section has shown that fractional differences are measured on the natural log scale. How then does one express variability in fractional terms? The standard deviation (SD) of a sample of n ($\geqslant 2$) points is the root mean square of their differences relative to the arithmetic mean. This suggests that log SDs and fractional SDs, like log differences and fractional differences, are equivalent. If so, following (3), the function

$$\mathrm{SD}(\ln x) - \frac{\mathrm{SD}(x)}{\mathrm{ML}(x)} \qquad (11)$$

ought to be small when ML(x) is suitably defined.

As with (3), (11) vanishes when ML(x) is defined as the logarithmic mean LM(x), with its definition generalized as follows:

$$\begin{aligned} \mathrm{LM}(x) &= \frac{\mathrm{SD}(x)}{\mathrm{SD}(\ln x)} \qquad && \mathrm{SD}(x) > 0 \\ &= \mathrm{AM}(x) && \mathrm{SD}(x) = 0 \end{aligned} \qquad (12)$$

This reduces to the previous definition (4) when $n = 2$. Thus based on the logarithmic mean the log SD and the fractional SD are identical by definition. The logarithmic mean, a ratio of two standard deviations, is clearly a remarkably inefficient measure of location, but as before the question arises, for $n > 2$ how similar is it to simpler measures of location? The answer is that it depends on the distributional form of x.

3.2.1. Log Normally distributed data. With the arithmetic mean in (11) the fractional SD is SD(x)/AM the coefficient of variation CV(x). (The suffix (x) for each mean is now dropped.) For log Normally distributed data, that is, where log x is Normally distributed, there is an exact relationship between CV(x) and SD(ln x) [13] given by

$$\exp(\mathrm{SD}^2(\ln x)) = 1 + \mathrm{CV}^2(x)$$

Expanding the left side as the first three terms in a Taylor series and rearranging shows that $\mathrm{LM} \approx \mathrm{AM}\sqrt{\{1 + \mathrm{SD}^2(\ln x)\}/2}$, so that LM > AM > GM in expectation. This differs from the case when $n = 2$.

3.2.2. Transformed Normally distributed data. Generalizing to data where x^p is Normally distributed (for $x > 0$), the Box–Cox transformation [11]

$$\begin{aligned} y &= \frac{x^p - 1}{p\,\mathrm{GM}^{p-1}} \qquad && p \neq 0 \\ &= \mathrm{GM}\ln(x) && p = 0 \end{aligned}$$

provides a direct comparison of the logarithmic mean and the geometric mean. The maximum likelihood estimate of the power p minimizes SD($y|p$), so the SDs of different power transformations of x can be compared, particularly the cases $p = 1$ and $p = 0$:

$$\frac{\mathrm{SD}(y|1)}{\mathrm{SD}(y|0)} = \frac{\mathrm{SD}(x)}{\mathrm{GM\,SD}(\ln x)} = \frac{\mathrm{LM}}{\mathrm{GM}}$$

In general, LM and GM are different. If x has a log Normal distribution, GMSD(ln x) < SD(x) and GM < LM, as was shown above, while if x is Normally distributed GMSD(ln x) > SD(x)

Table II. Expected values of the arithmetic, logarithmic and geometric mean for gamma distributions with different parameters α.

Arithmetic mean α	Logarithmic mean $\sqrt{\left(\frac{\alpha}{\text{trigamma}(\alpha)}\right)}$	Geometric mean exp(digamma(α))
1	0.78	0.56
4	3.75	3.51
9	8.75	8.50
16	15.75	15.50
25	24.75	24.50
49	48.75	48.50
64	63.75	63.50
100	99.75	99.50

and GM > LM. By symmetry the two means are likely to be the same when p is midway between 0 and 1, that is, the square root transform. This differs slightly from the cube root transform of Section 3.1.

The comparison of SDs under different Box–Cox power transformations applies to linear models with general design matrices [11], so that the equivalence of log SDs and fractional SDs extends naturally to confidence intervals and analysis of variance.

3.2.3. Gamma distributed data. The gamma distribution is another obvious distribution to consider, because of its link to the 1/3 power transformation. Wilson and Hilferty [14] showed that a χ^2 variate (which is also a gamma variate) raised to the 1/3 power is approximately Normally distributed. Thus the cube root mean applied to gamma distributed data is an approximately unbiased estimate of the median.

For a sample x from a gamma distribution with parameter α, the mean of $\log x$ is given by the digamma function $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$, and its variance by the trigamma function $\psi'(\alpha)$ [15]. The arithmetic mean of x is α, the geometric mean is exp(digamma(α)), and the logarithmic mean is $\sqrt{(\alpha/\text{trigamma}(\alpha))}$. Table II gives the three means for a range of α values, and shows that to a close approximation AM = LM + 0.25 = GM + 0.5. Thus LM $\approx$ (AM + GM)/2 for all $\alpha > 4$.

Simulation was used to establish the sampling properties of the different means; 400 samples of size 400 were drawn from a gamma distribution with parameter 25, and AM, GM, WM, CRM and LM were calculated for each sample along with the median. Parameter 25 corresponds to an arithmetic mean and variance of 25, so the SD is 5, the standard error (SE) for each sample 0.25 and the CV 0.2. Table III gives the arithmetic mean and SD across samples for the different measures of location, ranked by size. AM and GM are the largest and smallest, respectively, while WM and CRM are very close to the median, as predicted by Wilson and Hilferty [14]. The SDs for these four means are close to the expected SE of 0.25, while the median and particularly LM are less efficient. LM is, as shown above, midway between AM and GM.

In summary, the logarithmic mean is closely linked to the arithmetic and geometric means in a way that depends on the underlying frequency distribution. For distributions varying in skewness between the Normal and log-Normal, where the arithmetic mean always exceeds the geometric mean, the logarithmic mean shifts from above the arithmetic mean for log Normal data

Table III. Summary statistics across samples of measures of location for 400 random samples of size 400 drawn from a gamma distribution with parameter 25.

Measure of location	Arithmetic mean	Standard deviation
Arithmetic mean	24.99	0.249
Logarithmic mean	24.75	0.390
Median	24.66	0.313
Cube root mean	24.66	0.247
Weighted mean	24.66	0.247
Geometric mean	24.50	0.247

to below the geometric mean for Normal data. It equals the geometric mean for square root transformed data, and for gamma (or cube root) distributed data is exactly halfway between the arithmetic and geometric means.

The link between the three means shows that the logarithmic mean is a reasonable measure of location. This in turn confirms that the logarithmic SD in (11) is a form of fractional SD.

3.3. *Additivity and fractional regression coefficients*

The additivity property of percentages referred to in Section 2.2 requires that a sum of fractional differences is equal to the overall fractional difference. Log differences have this property, but conventional fractional differences do not. To illustrate this take three x values (x_1, x_2 and x_3) and calculate the fractional and log differences between pairs of them. Clearly

$$(\ln x_3 - \ln x_1) = (\ln x_3 - \ln x_2) + (\ln x_2 - \ln x_1)$$

whatever the values of x, indicating additivity, whereas in general

$$\frac{x_3 - x_1}{\mathrm{ML}(x_1, x_3)} \neq \frac{x_3 - x_2}{\mathrm{ML}(x_2, x_3)} + \frac{x_2 - x_1}{\mathrm{ML}(x_1, x_2)}$$

unless some of the x values are the same, or ML = LM.

A regression coefficient is a special sort of difference where additivity is assumed. Take the regression of x on some covariate t. The regression coefficient $\mathrm{B}(x|t)$ is the expected difference in x associated with a unit difference in t. Multiplied by t it gives the expected overall difference in x associated with t, the sum of the unit differences. By analogy with (3) and (11) this coefficient, expressed as a fraction of some measure of location ML of x, ought to be the same as the regression coefficient of ln x on t.

The function

$$\mathrm{B}(\ln x|t) - \frac{\mathrm{B}(x|t)}{\mathrm{ML}} \tag{13}$$

measures the disagreement between the two coefficients, where $\mathrm{B}(\ln x|t)$ is the regression coefficient of ln x on t. If $r(x, t)$ is the correlation between x and t then $\mathrm{B}(x|t) = r(x, t)\,[\mathrm{SD}(x)/\mathrm{SD}(t)]$, and correspondingly for B(ln $x|t$). Substituting for both terms in (13) and substituting LM (12) for ML gives:

$$r(\ln x, t)\frac{\mathrm{SD}(\ln x)}{\mathrm{SD}(t)} - r(x, t)\frac{\mathrm{SD}(x)}{\mathrm{SD}(t)\mathrm{LM}}$$

or

$$\frac{\mathrm{SD}(\ln x)}{\mathrm{SD}(t)}[r(\ln x, t) - r(x, t)]$$

To the extent that the two correlation coefficients are equal, the log regression coefficient is equal to the fractional regression coefficient based on the logarithmic mean. More generally, the log regression coefficient is seen to be a form of fractional regression coefficient, but uniquely among fractional coefficients it is also additive.

The results of the previous section also show that the log residual SD about the regression line is a form of fractional SD.

4. EXAMPLES

In this section the link between log differences and fractional differences is exploited to simplify the interpretation of results from analyses involving natural logs. There is no need to use the algebra of Section 3 – it is provided there only to confirm the link between natural log differences and fractional differences. The transform $y = 100\log_e x$ is used to give results in units of symmetric percentage differences, which are here called *sympercent* or s% for short.

4.1. Comparison of two numbers

In Section 2.2 the mean adult height of the two sexes was discussed, 177.3 cm for men and 163.6 cm for women. The symmetric percentage difference in height between them can now be derived as (100 log 177.3–100 log 163.6) or 8.044 sympercent. On average, men are 8.044 s% taller than women and women are 8.044 s% shorter than men.

4.2. Comparison of two group means

A common statistical procedure is the comparison of group means by Student's t-test. When the data are log transformed, the difference between means, the SE and the confidence interval of the difference are all in log units. This usually complicates the interpretation, but with the approach described here it becomes very simple.

Consider the confidence interval for the difference between two means. Bland and Altman [5, 7] compared bicep skinfold thickness in patients with Crohn's disease ($n = 20$) and coeliac disease ($n = 9$). The means for the two groups were 4.72 mm and 3.53 mm respectively, a mean difference of 1.18 mm with standard error 0.92 and a 95 per cent confidence interval of $(-0.71, +3.07)$ mm.

Bland and Altman repeated the analysis after natural log transformation, giving a difference between means of 0.296 (SE 0.205), with confidence interval $(-0.114, 0.706)$. Following

Table IV. Results for the multiple regression analysis of bone mineral content (BMC) in Gambian and English women [16]. The body size variables – BMC, bone width, weight and height – are analysed after 100 log transformation.

Term	Regression coefficient	Standard error	t	P
Country: Gambia	2.14	0.95	2.3	<0.05
100 log bone width	0.706	0.041	17.4	<0.001
100 log weight	0.205	0.028	7.5	<0.001
100 log height	0.371	0.111	3.3	<0.001
Age	4.04	0.42	9.5	<0.001
Age^2	-0.0875	0.0091	-9.6	<0.001
Age^3	5.0×10^{-4}	0.6×10^{-4}	8.3	<0.001
Intercept	-353.0	53.6		

Dependent variable: 100 log BMC

convention, they recommended antilogging the results to give a ratio of 1.34 with confidence interval (0.89, 2.03).

A simpler alternative is to use the 100 log$_e$ transformation. This gives a mean difference between groups of 29.6 (SE 20.5), with confidence interval ($-$ 11.4, 70.6), that is, the logged results times 100. However, now they are directly interpretable as sympercent differences. Bicep skinfold is on average 29.6 s% greater in the Crohn's than the coeliac group, with confidence interval ($-$ 11.4 s%, 70.6 s%). Equally bicep skinfold is 29.6 s% less in the coeliac than the Crohn's group, with confidence interval ($-$ 70.6 s%, + 11.4 s%). This is a simple, direct and entirely valid way of presenting the results for the log transformed data.

4.3. *Group comparison by regression analysis*

Another way to compare group means is by regression analysis, which is useful when adjusting for other covariates at the same time. The two groups are distinguished by a binary variable that takes the value 0 in one group (the baseline) and 1 in the other.

Prentice *et al.* [16] compared the bone mineral content (BMC) of the mid-shaft radius in Gambian and English women, adjusted for body size using multiple regression. Table IV shows the results of the analysis, with 100 log BMC as the dependent variable. The regression coefficient for The Gambia compared to England was 2.14 (SE 0.95) adjusted for bone width, weight, height and age, where the body size covariates were also 100 log transformed. This immediately shows that BMC was 2.14 s% greater in the Gambian women, or equivalently 2.14 s% less in the English women.

Table IV includes the results for the body size covariates in the analysis. They, like BMC, were 100 log transformed, and their coefficients, like the country coefficient, can be interpreted in units of BMC s%. The bone width coefficient of 0.706 (SE 0.041) for example shows that a 1 s% change in bone width was associated with a 0.706 s% change in BMC. (The same coefficient can also be interpreted as the power of bone width when the regression equation is antilogged, but this is not discussed further here.)

Another example appears in Table II of Cole *et al.* [17], where differences in height and weight between data sets estimated by regression analysis are presented in sympercent units.

4.4. Rate ratios

Rate ratios such as mortality ratios and odds ratios have hardly featured so far, but they fit into the same framework when transformed to natural logs. The log of a ratio is the same as a difference on the log scale (see Section 2.2). The regression coefficient in logistic regression is a log-odds ratio [18], and conventionally this is antilogged for presentation purposes, but the odds ratio is asymmetric in that reversing the sense of the comparison inverts it rather than changing its sign. Although users are familiar with this, there is still scope to present the odds as a 100 log odds ratio, which is a symmetric percentage difference in odds.

The same argument applies to the rate ratio. In a recent study comparing inequality in different European countries [19], the mortality rate ratio (RR) for all-cause mortality in England and Wales for manual as compared to non-manual workers was $\mathrm{RR}_o = 1.44$, indicating a 44 per cent excess mortality in manual workers. However, if the groups were switched, comparing non-manual to manual workers, the rate ratio would be $1/1.44 = 0.69$ and the non-manual mortality less by 31 per cent.

The two percentages 31 per cent and 44 per cent are very different, and there is no particular reason to prefer one over the other. The simpler symmetric alternative is to focus on 100 times the difference of the logs, or equivalently $100 \log \mathrm{RR}_o$, in this case 100 log 1.44 or 36 s%. This is 36 s% more manual mortality and equivalently 36 s% less non-manual mortality.

The same principle can be extended to compare several groups at once. The above study [19] had data for 11 European countries, and Figure 1 compares the levels of inequality in the different countries using England and Wales as the baseline. The percentage differences are calculated as $100 \log \mathrm{RR} \ - 100 \log \mathrm{RR}_o$, equivalent to $100 \log (\mathrm{RR}/\mathrm{RR}_o)$, and they show agreement within 10 s% across all the countries except France, where inequality is higher.

Spiegelhalter and Knill-Jones [20] have used the same form for the likelihood ratio (LR) in medical diagnosis, where they call $100 \log_e \mathrm{LR}$ the ‘weight of evidence’ for a particular symptom in favour of a particular diagnosis.

4.5. Standard deviations

Presenting summary statistics of logged data is a recurring problem. The log mean can obviously be antilogged to give the geometric mean, but what about the log SD? Conventionally the options are limited – either to tabulate the log SD as it stands, which is hard to interpret, or to antilog it and call it the geometric SD [1], or else to calculate centiles of the distribution assuming Normality on the log scale, and antilogging.

The approach proposed here offers two simpler alternatives. The log SD is equivalent to a CV, so it can be multiplied by 100 to put it in CV s% form, or alternatively it can be multiplied by the geometric mean to provide a form of SD in the original units.

Take the bicep skinfold data of Section 4.2 [5], where the log SDs for the two groups were 0.49 and 0.51, respectively. These multiplied by 100 are sympercent CVs of 49 s% and 51 s%, compared to the original CVs of 51 per cent and 56 per cent. The log-based results are smaller, which shows that the data are closer to log-Normal than Normal, as Bland and Altman pointed out [5].

Alternatively, a form of SD can be calculated as the product of GM and log SD. The geometric means were 4.20 mm for the Crohn’s patients and 3.12 mm for the coeliacs, which give SDs of 2.06 and 1.61 mm for the two groups, again smaller than the SDs of 2.42 and 1.96 mm on the original

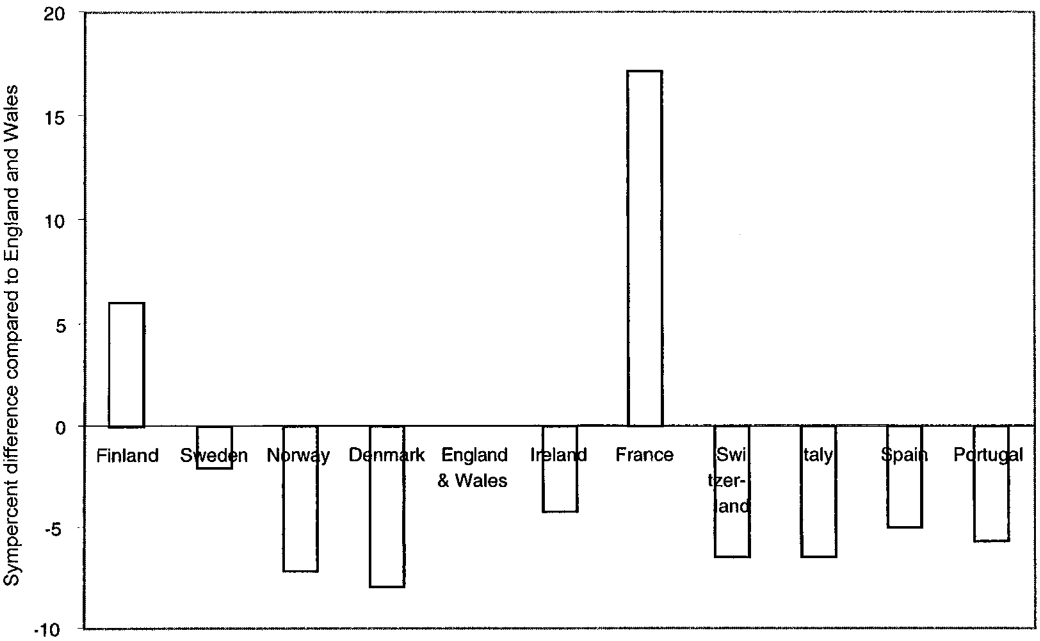

Figure 1. Excess manual versus non-manual all-cause mortality for 11 European countries [19] compared to England and Wales, expressed in sympercent units.

scale. The weakness of this approach is that because the SDs are in the original units of measurement, they lack the sympercent label to show how they were derived.

4.6. Analysis of variance

Fuller *et al.* [21] investigated inter- and intra-observer variability of height and weight in 12 subjects, each measured by six observers. The data were analysed by two-way analysis of variance, and variability for each measurement was expressed in two ways, as the residual standard deviation of the original measurements and the measurements transformed to 100 logs, that is, in sympercent units.

For weight, both analyses showed a highly significant observer effect (see Table V). The residual SDs were 0.0321 kg and 0.0451 s%. For comparison the residual CV was 0.0444 per cent , slightly less than the sympercent value, based on an arithmetic mean weight of 72.3 kg. This shows that the log transform for weight provided a slightly poorer fit to the data [11].

For height there were no significant differences between observers, and the residual SD was 0.705 cm or 0.405 s%. The residual CV was slightly greater than the sympercent value, 0.410 per cent based on an arithmetic mean height of 172.0 cm. In contrast to weight, the log transform for height fitted slightly better than untransformed height [11].

By doing the analysis on both scales the relative sizes of the observer effects can be compared, and the residual standard deviations are given directly in original and sympercent units.

Table V. Analysis of variance for weight in a study of inter-observer variability [21], in original and sympercent units.

Item	d.f.	Weight			100 log weight		
		SS	MS	F ratio	SS	MS	F ratio
Subject	5	9.91×10^3	9.01×10^2	8.75×10^5	2.00×10^4	1.82×10^3	8.93×10^5
Observer	11	0.167×10^{-1}	0.333×10^{-2}	3.24	0.334×10^{-1}	0.669×10^{-2}	3.28
Residual	55	0.567×10^{-1}	0.103×10^{-2}		0.112	0.204×10^{-2}	
Total	71	9.91×10^3			2.00×10^4		
Residual SD			0.0321 kg			0.0451 s%	

4.7. Regression coefficients

The examples so far have concentrated on the symmetry of sympercents. Regression analysis exploits their other important property, additivity. For a 100 log transformed dependent variable, each regression coefficient is in units of sympercents per unit of the independent variable. Multiplied by the value of the independent variable the coefficient gives the predicted total sympercent effect on the dependent variable, that is, the sum of the unit sympercent effects.

Neonatologists study fractional growth [22] – they measure the weight gain of premature babies in units of g/kg/d, that is g/1000g/d or ‰/d (since the units of weight cancel out). As a result, growth rate in these units can be estimated from the regression of 1000 log weight on age. Figure 2 shows daily weights in a premature baby between 10 and 38 days of age (weight gain takes a week or more to stabilize after birth), with the log regression line superimposed. The weight gain is given by the regression coefficient of 13.28 g/kg/d, and the residual standard deviation (RSD) is 10.8 g/kg. The g/kg units could alternatively be called sympermills, like sympercents.

For comparison, the regression of 1000 weight/(arithmetic mean weight) on age is also shown in Figure 2. It has the same regression coefficient, and a slightly larger RSD of 12.5 g/kg, indicating a higher correlation on the log than the linear scale.

As a second example, Lucas *et al.* [23] present the regression of log transformed insulin at 12 years on weight standard deviation score (SDS) in a group of children born premature. The combined model in their Table 1 gives the coefficient for weight SDS at 18 months as 0.12 or 12 s%, showing that a difference of 1 SDS unit in weight is associated with a 12 s% difference in insulin concentration.

5. DISCUSSION

The results have shown that a difference on the natural log scale is a form of fractional difference, using the logarithmic mean to derive the fractional difference. This is neither obvious nor well known, but it arises from the equivalence of $\delta(\ln x)$ and $\delta x/x$ in the limit (Section 3.1).

Under modest distributional assumptions, the logarithmic mean is closely related to the arithmetic and geometric means. The relationship of the three means under a gamma distribution, with the geometric mean half a unit less than the arithmetic mean and the logarithmic mean midway between them (Tables II and III), seems not to be well known.

The logarithmic mean is remarkably inefficient as an estimate of location, just 40 per cent efficient compared to the geometric mean (Table III). Even so it does crop up in practical

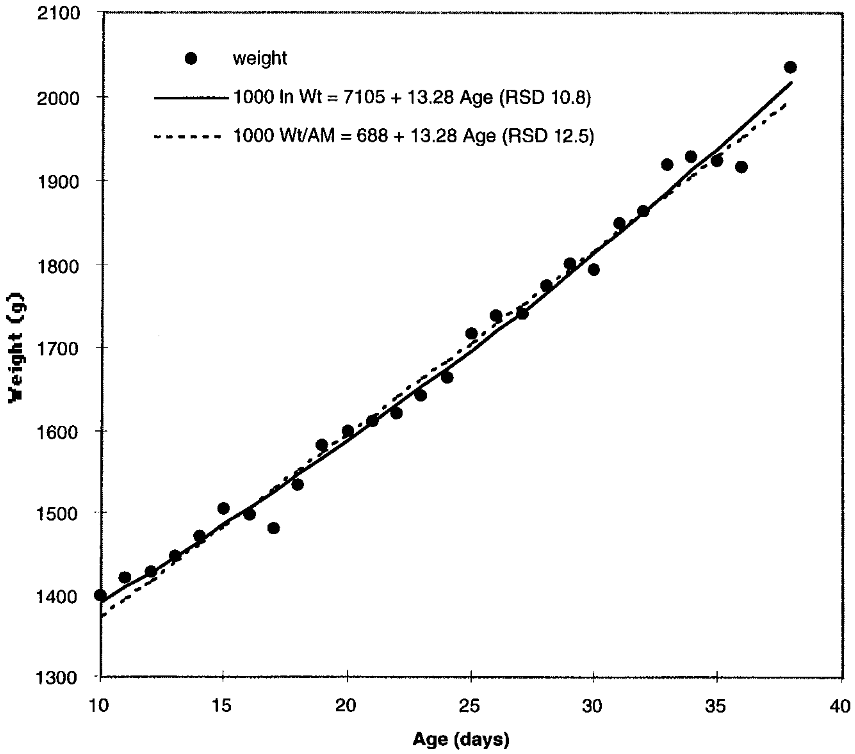

Figure 2. Fractional weight gain in a premature baby, in units of g/kg/d, estimated from the regression of (a) 1000 log weight and (b) 1000 weight divided by its arithmetic mean, on age.

applications. Coward *et al.* [24] for example showed that under certain assumptions the intake of water in growing animals over a period of time is equal to $\text{LM}(Q)\ln(C_1/C_2)$, where Q_1 and Q_2 are body water volumes and C_1 and C_2 are tracer concentrations at the start and end of the period.

The fact that the log difference is a fractional difference does not automatically make it the best or most 'natural' form of fractional difference. A leap of faith is required to see that it is optimal. The key argument in its favour is simplicity, as shown in three ways. First symmetry: it is directly analogous to the difference between two numbers, giving an answer in percentage rather than measurement units. Just as for the difference, swapping the numbers changes the sign of the percentage difference but not its value.

Secondly, the log difference is unique among fractional differences in being additive. This applies particularly in log regression.

Thirdly, there is only one way to calculate the log difference, while there are many different forms of fractional difference. The ordinary percentage difference has the first or second of the two numbers in the denominator, while the symmetric modification discussed here uses their mean. Yet there are many different means and they give different answers. Uniquely, the log difference gives the fractional difference without involving a denominator at all.

So the case is simple: percentage differences ought to be calculated as differences on the natural log scale. To this end, the transformation $y = 100\log_e x$ should be more widely recognized as giving results that are directly interpretable in symmetric percentage units. The same applies to standard deviations and regression coefficients.

The one drawback with this philosophy is that it can be confusing. However a clear statement in the Methods that 100 log differences are interpreted as symmetric percentage differences should avoid the problem. Ideally, though, a new nomenclature is needed, and 'sympercent' is proposed as an abbreviation for 'symmetric percentage', with the symbol 's%'.

Although natural logs are useful for deriving symmetric percentage differences there is a further issue: should sympercents also replace per cents for expressing change over time – should inflation rates, interest rates and other temporal rates be made symmetric and additive by basing them on log differences? Törnqvist *et al.* [8] felt that they should, yet their paper is little cited, which perhaps reflects the preparedness of econometricians for such a major change, despite the potential benefits in terms of symmetry and additivity. The proposal here is not so far-reaching and should be less controversial. The most obvious advantage of a shift to sympercents throughout would be less confusion, since the units would be consistent universally.

Gaddum [25] did a lot to publicize the log transformation, inventing the term 'lognormal' distribution and recommending the use of common logs. He called the standard deviation of $\log_{10}$ transformed data λ, and used it to classify different measurements in terms of their variability. Had he used the natural log transformation instead, his λ would have been effectively the coefficient of variation, rather than something 2.3 times smaller.

When should data be log transformed? The usual criteria are the presence of heteroscedasticity, skewness or non-linearity, and nothing here is intended to influence that decision – the paper is concerned only with the presentation of results once a log transformation has been chosen. However there are situations where a log transform may be valid even though the original data are not obviously skew. Height is an example of Normally distributed data, yet log height is also acceptably Normal as well – see Section 4.6 for example. Height has a small CV, 4 per cent or so [17], which means that transformation does not introduce much skewness. In any case, if results are required in percentage units, this may be sufficient reason for a log transformation even though the data are not particularly skew. This relates to Keene's argument [3] that the log transformation is special, and should be given equal status with analysis on the original scale.

In conclusion, a theoretical and practical case has been made for using the natural log scale, multiplied by 100, to calculate and present percentage differences, standard deviations, regression coefficients and rate ratios. The hope is that as a result, natural logs will appear less obscure and more accessible – in a word, natural.

ACKNOWLEDGEMENTS

It is largely due to the support and encouragement of Ann Prentice and Doug Altman that this paper, 25 years in the writing, has finally seen the light of day. Henry Daniels pointed me to the Wilson–Hilferty transformation, and Martin Bland, Ruth Morley, Rory Wolfe and two anonymous referees made useful suggestions. I am grateful to them all.

REFERENCES

1. Kirkwood TBL. Geometric means and measures of dispersion. *Biometrics* 1979; **35**:908–909.
2. Bland JM, Altman DG. Statistics notes 16. Logarithms. *British Medical Journal* 1996; **312**:700.
3. Keene ON. The log transformation is special. *Statistics in Medicine* 1995; **14**:811–819.
4. Altman DG. *Practical Statistics for Medical Research.* Chapman and Hall: London, 1991; pp 199–203.

5. Bland JM, Altman DG. Statistics notes 19. The use of transformation when comparing two means. *British Medical Journal* 1996; **312**:1153.
6. Armitage P, Berry B. *Statistical Methods in Medical Research*. 2nd edn. Blackwell: Oxford, 1987; pp 360–364.
7. Bland M. *An Introduction to Medical Statistics*. 2nd edn. Oxford Medical Publications: Oxford, 1995; pp 161–165.
8. Törnqvist L, Vartia P, Vartia YO. How should relative changes be measured? *American Statistician* 1985; **39**:43–46.
9. Freeman JV, Cole TJ, Chinn S, Jones PRM, White EM, Preece MA. Cross-sectional stature and weight reference curves for the UK, 1990. *Archives of Disease in Childhood* 1995; **73**:17–24.
10. Carlson BC. The logarithmic mean. *American Mathematical Monthly* 1972; **79**:615–618.
11. Box GEP, Cox DR. An analysis of transformations. *Journal of the Royal Statistical Society Series B* 1964; **26**:211–252.
12. Diewert WE. Superlative index numbers and consistency in aggregation. *Econometrica* 1978; **46**:883–900.
13. Johnson NL, Kotz S, Balakrishnan N. *Continuous Univariate Distributions. Volume* 1, 2nd edn. Wiley: New York, 1994.
14. Wilson EB, Hilferty MM. The distribution of chi-square. *Proceedings of the National Academy of Sciences*, *Washington* 1931; **17**:684–688.
15. Abramowitz M, Stegun IA. *Handbook of Mathematical Functions with Formulas*, *Graphs*, *and Mathematical Tables*. Wiley: New York, 1972.
16. Prentice A, Shaw J, Laskey MA, Cole TJ, Fraser DR. Bone mineral content of British and rural Gambian women aged 18–80+ years. *Bone and Mineral* 1991; **12**:201–214.
17. Cole TJ, Freeman JV, Preece MA. British 1990 growth reference centiles for weight, height, body mass index and head circumference fitted by maximum penalized likelihood. *Statistics in Medicine* 1998; **17**:407–429.
18. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. Wiley: New York, 1989.
19. Kunst AE, Groenhof F, Mackenbach JP, Group EW. Occupational class and cause specific mortality in middle aged men in 11 European countries: comparison of population based studies. *British Medical Journal* 1998; **316**:1636–1642.
20. Spiegelhalter DJ, Knill-Jones RP. Statistical and knowledge-based approaches to clinical decision-support systems, with an application to gastroenterology (with Discussion). *Journal of the Royal Statistical Society Series A* 1984; **147**:35–76.
21. Fuller NJ, Jebb SA, Goldberg GR *et al*. Inter-observer variability in the measurement of body composition. *European Journal of Clinical Nutrition* 1991; **45**:43–49.
22. Lucas A, Gore SM, Cole TJ *et al*. Multicentre trial on feeding low birth weight infants – effects of diet on early growth. *Archives of Disease in Childhood* 1984; **59**:722–730.
23. Lucas A, Fewtrell M, Cole TJ. Fetal origins of adult disease–the hypothesis revisited. *British Medical Journal* 1999; **319**:245–249.
24. Coward WA, Cole TJ, Gerber H, Roberts SB, Fleet I. Water turnover and the measurement of milk intake. *Pflugers Archiv European Journal of Physiology* 1982; **393**:344–347.
25. Gaddum JH. Lognormal distributions. *Nature* 1945; **156**:463–466.