Taking logs - why and how?

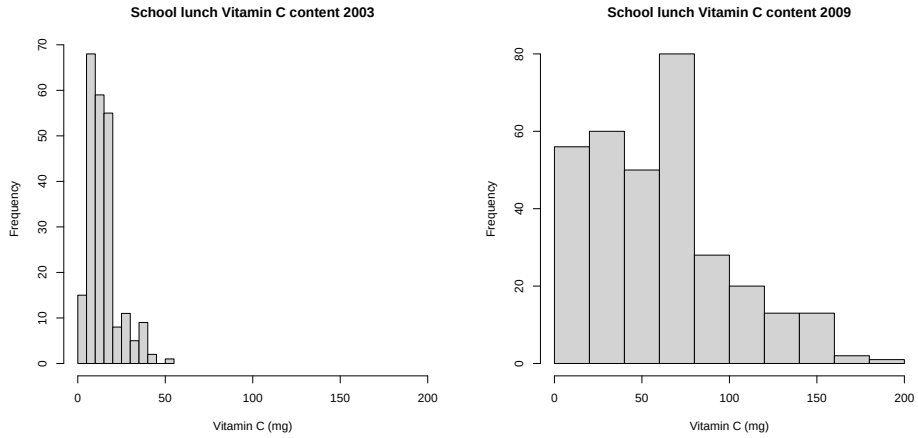
J N S Matthews
Biostatistics Research Group,
Newcastle University

1 An example

Consider the data shown in Figure 1. They are from a study of nutrient intakes in six schools in Northumberland, surveyed in 2003 and again in 2009 (Spence et al., 2013). The variable shown is the average amount of vitamin C (in mg) in school (not packed) lunches over three consecutive days for children in these two years. In the paper these data were part of an analysis which used a random effects regression analysis, so that numerous other covariates could be accommodated. However, for illustrative purposes we can focus on a simple comparison between the two years. Summary statistics for the two years are in Table 1.

Year	n	Mean	SD	Min	Q1	Median	Q3	Max
2003	233	14.5	8.6	0.1	8.5	12.5	17.4	52.4
2009	323	60.0	38.4	5.8	26.2	59.1	77.8	184.7

Table 1: Summary statistics for the vitamin C intakes (mg)



(a) Sample in 2003 ($n = 233$) (b) Sample in 2009 ($n = 323$)

Figure 1: Raw data on vitamin C content of school lunches (Spence et al., 2013)

When comparing two groups, initial thoughts probably turn to a *t*-test, but anyone faced with comparing the vitamin C intakes between these two years would likely pause before doing so. The *t*-test assumes two Normally distributed groups, with a common population standard deviation (SD), and this seems far from the case here. The distributions in Figure 1 look skewed and quite far from Normal, and each SD is more than half the corresponding mean (so that an assumption of Normality would imply a reasonable proportion of negative vitamin C values in the population). Moreover, the means (and medians, in this example) are very different from the mid-point of the distribution, as based on the minimum and maximum values.

2 Approaches to comparing the years

2.1 Distribution-free or non-parametric methods

These methods are often the first port of call for those faced with data which do not appear Normal. The methods are usually based on ranks and is a route that is, perhaps, taken more readily by non-specialists. One of the reasons for may be that it will often be thought that because the methods do not assume any specific distribution, they are assumption-free. However, this

is not the case, especially when it comes to estimation. While the methods do have their uses, they should be used only after considering the following points.

1. The usual methods, such as Mann-Whitney, assume that the data in group 1 follow a distribution $F_1(x)$ and in group 2, $F_2(x)$, and the hypothesis associated with the usual test is $\mathrm{H}_0 : F_1(\cdot) = F_2(\cdot)$. So it tests that the samples come from the same population. The conventional unpaired *t*-test, with population SDs assumed equal, does the same, with $F = \Phi$. However, variants of the *t*-test which do not assume equal SDs assess common location of the populations, notwithstanding possible differences in dispersion. Of course, it may well be that there is limited value in a test of location, while allowing other aspects of the population to differ.

2. These tests are based on ranks - so having put great effort into measuring some variable, its precise value is ignored in favour of its rank relative to other values. This might be a strength, insofar as it allows data arising only from ordinal categories to be analysed, and in other cases provides some insensitivity to large values. With rank-based methods, whether not analysing the observed values directly is desirable in a given application is worth considering.

3. As we so often incant, hypothesis testing is only a small part of the analysis of data, and estimation, especially interval estimation, is always emphasised. Estimation when using these methods often focuses on the medians in the two groups, say θ_1, and θ_2. The use of medians for data such as that in Figure 1 is often advised as medians are less affected than means by some unusually large values in the sample. In terms of the theory of robust estimators (Huber, 1981), the median has a *high breakdown* point of 50%, i.e. up to 50% of the values in a sample can be changed without changing the value of the median. This might strike some as a large proportion, suggesting that the median is perhaps a little too robust.

4. Estimates of the median, or median difference are seldom accompanied by standard errors but this is not too troublesome as interval estimators are readily available. However, most methods in common practice are based on the assumption that one population is simply a translation

of the other, i.e. $F_2(x) = F_1(x - \Delta)$, where Δ is, say, a difference in medians. The assumption that the two populations share the same dispersion applies, so the idea the method is assumption free is not right. Moreover, producing an interval estimate for the parameter Δ makes the term non-parametrics slightly irritating.

5. A comprehensive examination of a large dataset will seldom comprise simple group comparisons, and often analyses taking simultaneous account of many variables need to be presented. Most of the commonly used distribution-free methods are not rich enough for the more elaborate analyses that are commonplace in biostatistics. Sometimes you see simple, preliminary analyses done using distribution-free methods, but these are followed by the fitting of a normal-theory model to the same data, because a suitably sophisticated distribution-free is not readily available. This can result in an awkward and unconvincing presentation.

2.2 Just ignore the skewness

One could just analyse the data as it has been collected and ignore the technical problems the skewness poses. Certainly, for samples of reasonable size, such as those in Figure 1, an analysis based on means may be entirely respectable, given the close approximation to Normality conferred on the sample means by the Central Limit Theorem. The difference in sample SDs is a mite troubling[1], but as just remarked, the assumption of equal population dispersion is also present in a distribution-free analysis. At least with a *t*-test you could use a version which does not make this assumption.

2.3 Logs: why should they be used?

Of course, most statisticians would not use either approach outlined in Section 2 - they would take logs of the data. The reason usually adduced for this is that the logs of skewed data, such as in Figure 1, will be closer to Normal than the untransformed data. This is often seen by collaborators as a baffling device statisticians use in the quiet of their offices to maintain their purity, and is a price that has to be paid for having professional statistical help.

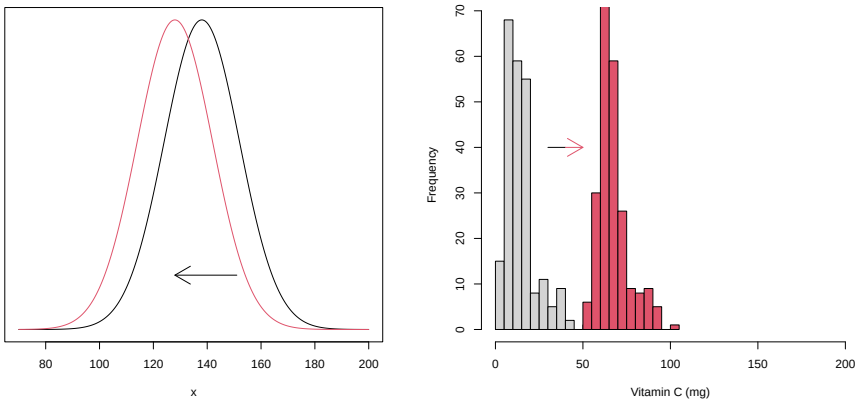
[1]But note that it would not pose a problem if the sample sizes were approximately equal

Indeed, statisticians often emphasise, in courses to non-specialist colleagues, the moral turpitude of analysing skewed data using Normal-theory methods. As most non-statisticians are less than confident with logs, this probably explains the frequent use of distribution-free methods. After concerns over sample size, perhaps the second most common issue confronting consulting statisticians is the cry that "my data aren't Normal" (or, unforgivably, "my data are non-parametric").

However, perhaps the real benefit of taking logs is because of the inadequacy of the interpretation of the methods in Section 2 when applied to skewed, positive variables. This is something that should be equally concerning, and equally apparent to statistician and collaborator alike. The description of the difference between the groups from either approach presented in Section 2 is in terms of a confidence interval for a difference in means or medians, i.e. it is envisaged that one population is simply a *translation* of the other. While this may be entirely plausible for two groups of systolic blood pressure, with means typically around 130 mmHg (with SD of 18 mmHg), it is very unconvincing for the vitamin C data: see Figure 2.

The difference in mean, or median, between the 2003 and 2009 data is about 50mg, and the mean in 2009 is about four times that in 2003. However, the 2003 data shifted by 50mg, shown in Figure 2b, looks nothing like the 2009 data shown in Figure 1b. In the shifted data the dispersion has not changed from 2003, whereas the SD in 2009 is much larger than in 2003 (see Table 1), and the shape of the shifted distribution, with a gap of over five SDs between the minimum value and zero, is quite different. Indeed, it often seems that skewed positive clinical variables have many values 'close' to zero with a reasonable positive tail - to put it very crudely. It is as if the positivity constraint is what is driving the skewness.

So, the simple shift, implicit in the analyses in Section 2, gives a very poor description of the difference between the 2003 and 2009 vitamin C values. What might be a better description? Suppose that the 2003 vitamin C observations are denoted by $x_i, i = 1, \ldots, n = 233$, Note also that the mean of the 2009 values is about four times that of the 2003 values. Consider, purely for illustration, the values $u_i = x_i f_i$, where the f_i are independent realisations from a Gamma distribution with mean 4 and variance 1. Histograms of the scaled values, u_i, and the 2009 sample are are shown in Figure 3.



(a) Shifted systolic BPs (b) Vitamin C shifted by c. 50mg

Figure 2: Plots of shifted hypothetical populations of systolic BP and shifted 2003 vitamin C data.

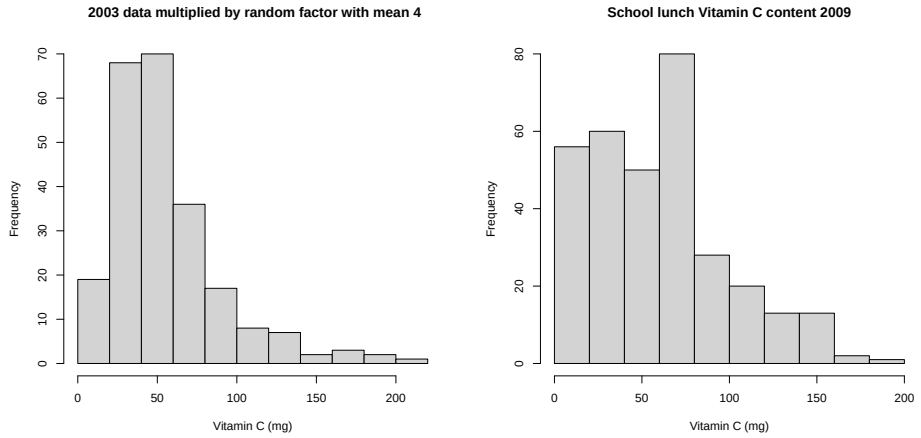
While the two histograms are far from identical, the scaled values in Figure 3a are much more similar to the 2009 observations than the shifted values in Figure 2b. A better description of the change in vitamin C content from 2003 to 2009 is likely to be found from a multiplicative change rather than an additive one. As standard statistical techniques deal in additive changes, perhaps the principal contribution of the log transformation is that it allows multiplicative changes to become additive ones.

3 Using the log transformation

The usual approach to using a transformation in a statistical analysis using Normal-theory methods is as follows.

1. Select a suitable transformation, $g(\cdot)$, so that the $g(x_i)$ have a distribution that can reasonably be assumed to be Normal: here $x_i, i = 1, \ldots, n$ are the observed data.

2. Analyse the $g(x_i)$ using standard methods.



(a) Vitamin C: scaled 2003 values (b) Vitamin C: 2009 values

Figure 3: Vitamin C values from 2003 scaled by Gamma variables and observed values from 2009

3. Present the results of the analysis *on the scale of the original observations*, usually by appropriate use of the inverse transformation, $g^{-1}(\cdot)$.

Most of the time statisticians emphasise point 1, with $g(.)$ chosen so that the transformed data have desirable properties: this is the main aim presented in Box and Cox (1964). However, if point 3 cannot be accomplished in a way that is compelling and comprehensible, then the whole exercise will be much less convincing.

3.1 Taking logs of the vitamin C data

Suppose that the observations on vitamin C in 2003 are $x_1, \ldots, x_{n_3}$, with $n_3 = 233$ and for 2009 they are $y_1, \ldots, y_{n_9}$, with $n_9 = 323$. If the analysis uses the logged values, then the main summary will be the arithmetic means of the logged values, namely[2]

$$m_3 = \frac{1}{n_3} \sum_{i=1}^{n_3} \log x_i$$

[2]These are natural logs but in practical terms the base is immaterial

and

$$m_9 = \frac{1}{n_9} \sum_{i=1}^{n_9} \log y_i.$$

Our earliest encounters with means, or averages, would be that m_3 is a 'typical' value for vitamin C values, but on the log scale. It is then quite natural to compute $\exp(m_3)$ in order to retrieve a 'typical' value on the familiar scale on which the x_i were measured. Thus $\exp(m_3)$ and $\exp(m_9)$ provide plausible summaries of location for the two groups.
Perhaps the most important use of m_3 and m_9 is to summarise the difference between 2003 and 2009. To this end it is natural to compute $m_9 - m_3$, i.e. the difference on the scale with a Normal distribution, here the log scale. Is it sensible to anti-log this quantity, i.e. $\exp(m_9 - m_3)$, to get back to a familiar scale? How does this relate to $\exp(m_3)$ and $\exp(m_9)$? For the log transformation, this step provides one of its most valuable properties, namely

$$\exp(m_9 - m_3) = \frac{\exp(m_9)}{\exp(m_3)}.$$

So the antilog of the difference in means is the *ratio* of the anti-logged m_3 and m_9. This has two virtues

1. The anti-log of the principal measure of the difference between years is determined by the individual means.

2. The difference between the years is naturally quantified by a ratio, i.e. a *multiplicative* difference.

Point 1 is important because it is really only point estimates where it is intuitively clear that applying $g^{-1}(\cdot)$ is appropriate.
Describing differences in terms of a ratio - such as "mean vitamin C levels were 3.9 times higher in 2009 than 2003" - is simple, understandable and is often very attractive to collaborators.

3.2 The geometric mean

3.2.1 Definition and elementary properties

In the previous subsection it was indicated that means are related multiplicatively but this implicitly refers to $\exp(m_3)$ and $\exp(m_9)$ as *means*. This is

correct but they are not arithmetic means. Note that, for example,

$$\exp(m_3) = \exp\left(\frac{1}{n_3}\sum_{i=1}^{n_3}\log x_i\right) = \sqrt[n_3]{\left(\prod_{i=1}^{n_3} x_i\right)},$$

which is the *geometric* mean of the observed vitamin C values. Note that the geometric mean is only defined for strictly positive values.

1. The geometric mean is a widely studied quantity, which has a similar elementary motivation to the arithmetic mean - the values are combined by multiplying rather than adding. If the geometric mean of n elements is G_n and the arithmetic mean is A_n, then

   $$G_n \leq A_n.$$

   with equality only if all elements of the sample are equal.

2. With positively skewed data, the median is less than the arithmetic mean and, usually, the geometric mean is closer to the median than the arithmetic mean.

3. Large values perturb the geometric mean less than the arithmetic mean because the log transformation reduces the influence of large observations.

4. As such, for skewed data, the geometric mean can be seen as a useful compromise between the excessive sensitivity of the arithmetic mean and the overly robust median.

5. The geometric mean can be sensitive to changes in small values - see later.

3.2.2 Some theoretical considerations

The log transformation works well with various skewed distributions of the sort shown in Figure 1, but theoretical matters are most readily developed if the skewed data are assumed to have a log-Normal distribution. A positive random variable, Y, has a log-Normal distribution if $Y = \exp(X)$, where X is Normally distributed with mean μ and variance σ^2.

The following properties are worth noting. In doing so, it is helpful to recall that the moment generating function (MGF) of a Normal variable is

$$M(t) = \mathsf{E}[\exp(tX)] = \exp(\mu t + \tfrac{1}{2}t^2\sigma^2).$$

1. As $Y = \exp(X)$, $\mathsf{E}[Y] = M(1) = \exp(\mu + \frac{1}{2}\sigma^2)$. So the arithmetic mean of Y is larger than $\exp(\mu)$.

2. As exp is monotone increasing, $\frac{1}{2} = \Pr(X < \mu) = \Pr(Y < e^{\mu})$, so $\exp(\mu)$ is the median of Y.

3. The variance of Y is $M(2) - M(1)^2 = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1) = \mathsf{E}[Y]^2(\exp(\sigma^2) - 1)$. So the variance of Y depends on μ and, moreover, is proportional to the square of its mean. Note that the coefficient of variation, i.e. the SD divided by the mean, is $\sqrt{e^{\sigma^2} - 1}$: and for small σ, this is approximately σ.

4. The geometric mean of a positive random variable, Y, is defined as $\exp(\mathsf{E}[\log Y])$. For log-Normal Y, this is $\exp(\mu)$, i.e. in this case the geometric mean coincides with the median.

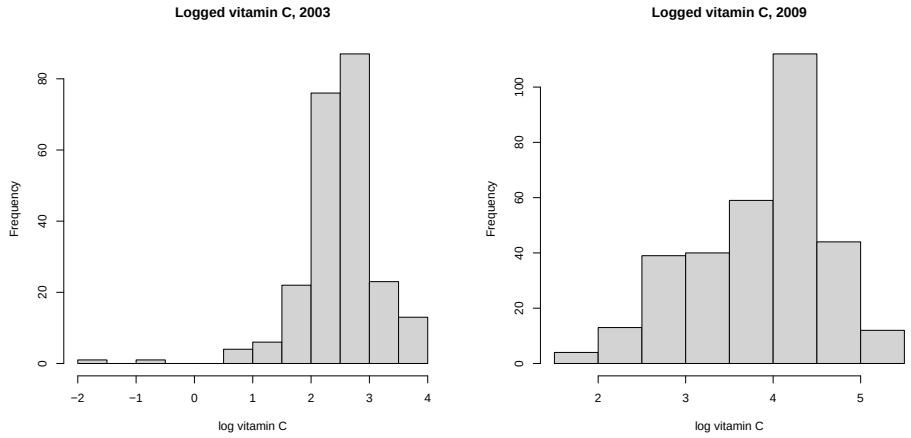
5. Note that, regardless of the distribution of Y, its geometric mean is always less than $\mathsf{E}(Y)$, i.e. the arithmetic-geometric mean inequality carries over to random variables. This follows by noting that log is concave and applying Jensen's inequality (Chung, 1974, p.47).

4 Application of logs to vitamin C data

The logs of the vitamin C values are shown in Figure 4 and are closer to Normal than the original distributions. Summary statistics for the logged values are in Table 2.

Year	n	Mean	SD	Mean (logs)	SD (logs)	Geometric mean
2003	233	14.5	8.6	2.491	0.680	12.1
2009	323	60.0	38.4	3.853	0.752	47.2

Table 2: Arithmetic and geometric means and associated quantities for vitamin C intakes: means and SD in mg



(a) log vitamin C values for 2003 (b) log vitamin C values for 2009

Figure 4: Histograms of the log (base e) vitamin C for 2003 and 2009

The geometric means in Table 2 are indeed smaller than the arithmetic means and, indeed, smaller than the medians shown in Table 1. The ratio of the SDs on the log scale is much closer to one than that for the SDs on the original scale.

4.1 Main comparison

The principal comparison is now made through

$$\exp(3.853 - 2.491) = \exp(1.362) = 3.90 = \frac{\exp(3.853)}{\exp(2.491)},$$

that is, the geometric mean vitamin C intake in 2009 was about 3.9 times what it was in 2003.

4.2 Measure of uncertainty

This is where some care is needed. The basic notion is that while it is sensible to apply $g^{-1}(\cdot)$ to point estimates of $g(\cdot)$-transformed quantities, it

is far from clear that applying $g^{-1}(\cdot)$ to measures of spread or difference will yield anything meaningful.
The analysis has been carried out on the log-scale and on that scale the uncertainty in the difference in means is readily presented in terms of a confidence interval. So, in the usual way, a 95% confidence interval is

$$d_L, d_U = (m_9 - m_3) \pm 1.96 \times 0.723 \sqrt{\left(\frac{1}{233} + \frac{1}{323}\right)} = (1.241, 1.485)$$

As d_L and d_U are estimates of the 2.5% and 97.5% points of the sampling distribution of the difference in sample means, it is legitimate to anti-log these to get the corresponding points for the sampling distribution of the ratio of geometric means, namely $\exp(1.241) = 3.46$ and $\exp(1.485) = 4.41$. Thus the results of the analysis can be more fully given by indicating that the geometric mean vitamin C intake in 2009 is 3.90 times that in 2003, and the 95% confidence interval for the factor 3.90 is (3.46,4.41). Note that the point estimate is less than the mid-point of the confidence interval: such asymmetric interval estimates are to be expected for skewed distributions.

4.3 Hypothesis test

The test of the hypothesis that $\mu_3 = \mu_9$, i.e. the equality of the arithmetic means on the log scale is the same, assuming a log-Normal distribution, as the test of the equality of the geometric means, $e^{\mu_3} = e^{\mu_9}$. So the *t*- and *p*-values to be quoted are those from the *t*-test on the log scale, which for the vitamin C data are 17.8 and $p < 10^{-15}$. Note that testing the equality of the arithmetic means on the log scale does not correspond to testing the equality of the arithmetic means on the original scale, unless the variances in the two groups are taken to be the same.

4.4 Does anti-logging the sample SD make sense?

While anti-logging the m_j yields an interpretable and useful quantity, there is little value in presenting $\exp(s)$, where s is the sample SD of the logged values. It is certainly not the SD of the unlogged data. For skewed data the spread is usually related to the mean - Table 1 shows that the mean *and* SD are notably larger for 2009. As such the variance of a skewed variable will

usually be partially determined by measures of location as well as spread. Thinking in terms of log-Normal observations, the variance is dependent on μ, i.e. $\exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$. For this distribution, μ cannot feature in the distribution of s, so $\exp(s)$ cannot be directly interpreted in terms of dispersion on the unlogged scale[3]. Measurement of uncertainty is best presented in terms of interval estimates, with the standard error playing no explicit role.

5 Some miscellaneous comments

5.1 Variance stabilising transformations

The analysis of the results is much easier if the variation in the data is largely unrelated to the mean. We add ϵs at the end of our models and give them zero means and *constant* variances. However, real data, especially skewed data, will often have variation that depends on the mean level of the response. This is quite natural - a quantity may have a constant percentage variation, which will translate to varying additive variances in groups with different mean responses. To find a transformation such that on the new scale the variance is largely constant can be done using the delta-method. Suppose X is a (positive) random variable and $f(\cdot)$ is a transformation such that $\mathsf{var}(f(X)) = \sigma^2$ (not a function of μ), then

$$\mathsf{var}(f(X)) \approx \mathsf{var}(f(\mu) + (X - \mu)f'(\mu)) = [f'(\mu)]^2\mathsf{var}(X) = \sigma^2.$$

If $\mathsf{var}(X) = k\mu^2$ (i.e. the SD of the response is proportional to μ), then we need $f'(\mu)\mu$ to be independent of μ, i.e. $f(\cdot) = \log(\cdot)$.
If the mean and variance are related, then it is perhaps natural for the variation to be proportional to the mean on the same scale as the observations, i.e. the dependence is such that it is the SD, not the variance, that is proportional to the mean. This leads to the log being the transformation which stabilises the variance - something reflected in Table 2.

[3]As seen above, $\exp(s^2)$ is related to the coefficient of variation.

5.2 Effect of log transforming on positive random variables

Suppose that X is a positive random variable with mean μ and SD σ. If the mean is similar to the SD, say σ/μ is up to, say, 2, then the distribution of X is likely to be quite skewed, and a log transformation will be helpful. If the CV, σ/μ, is small then writing $X = \mu + \sigma U$, so U has zero mean and SD one, then

$$\log(X) = \log(\mu + \sigma U) = \log(\mu) + \log(1 + \frac{\sigma}{\mu}U) \approx \log(\mu) + \frac{\sigma}{\mu}U.$$

Consequently, $\log(X)$ is a linear transformation of X, with approximate mean $\log \mu$ and SD σ/μ, i.e. the CV of X is approximately the SD of $\log(X)$.
In practical terms this suggests that if a variable is noticeably skewed (large CV), then logging it will be helpful, whereas for smaller CVs, logging is essentially a linear transformation that is not harmful. So, if in doubt - take logs.

5.3 Other transformations

A common alternative to the log transformation is a power transformation - as used in the Box-Cox family of transformations (Box and Cox, 1964). If X^p is closer to Normal than X, then it is natural to describe differences between groups by differences in means on the X^p scale. A snag is that back-transforming to the original scale is more awkward.
As an example, consider the square root transformation, i.e. $p = \frac{1}{2}$ - a transformation which may be used if the *variance* is proportional to the mean. Then the difference between the means would be

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{x_i} - \frac{1}{m}\sum_{i=1}^{m}\sqrt{y_i} = M_x - M_y, \text{say.} \quad (1)$$

The summaries M_x, M_y are on the root scale - if the observations are weights, they are in $\sqrt{\text{kg}}$. While M_x^2, M_y^2 are sensible summaries on the original scale, squaring (1) gives $M_x^2 + M_y^2 - 2M_xM_y$, which is not really interpretable.
There is, therefore little to be done other than to consider $M_x^2 - M_y^2$. But this does not acknowledge that differences are best taken on the transformed scale, and that differences on that scale may correspond to some other measure of discrepancy when back-transformed - as was the case with the log transformation, which led to multiplicative effects on the original scale.

5.4 Inference about the arithmetic mean

For most skewed data, the geometric mean is an appropriate summary. However, there are some cases where the arithmetic mean remains the most pertinent summary. One such is cost data, which are often skewed. The arithmetic mean, A, of the cost of a specified operation is useful because nA is the expected cost of performing n such operations. A log-Normal distribution could still be used - the logged costs would provide estimates of μ and σ^2 - but care would be needed to make inferences about the arithmetic mean, through $\exp(\mu + \frac{1}{2}\sigma^2)$, rather than use the geometric mean. This is a specialist area and the log-Normal may not be the best choice - some relevant papers are Thompson and Nixon (2005); Ng et al. (2016).

5.5 Zeroes in the data

This article has considered only skewed positive data. However, from time to time applied statisticians are faced with skewed data that contains a few zeroes. For those persuaded that a log transformation is needed for skewed data, these are the bane of their life, as the logarithm of zero is undefined.

5.5.1 Lots of zeroes

If the proportion of zeroes is substantial (and 'substantial' will need to be judged in context - generally more than a few percent of zeroes is very troublesome), then this is a clearly a germane feature of the data and needs to be addressed in any model used for the data. A very simple approach might be to reduce the outcome to 0 or 1, where 1 simply means 'not zero' and conduct a logistic regression. This analysis would then be supplemented with an analysis of the non-zero outcomes, which can be logged. This might suffice but more sophisticated analyses are possible, often using this general approach but linking the two models in some way. Data of this kind are sometimes referred to as 'semi-continuous' and Su et al. (2009) provides an interesting introduction.

5.5.2 Not too many zeroes

Various *ad hoc* approaches may be profitable when the number of zeroes is small relative to the sample size. A simple device is simply to add a 'small' positive quantity to each observed value - i.e. analyse not $\log y$ but $\log(y+\epsilon)$.

Of course, what constitutes 'small' needs to be judged in the context of the data. It may also be prudent to undertake sensitivity analyses - judging the effect on pertinent inferences of varying ϵ. This may not turn out to be as good as might be hoped - zeroes will be replaced in the analysis by $\log \epsilon$, and as $\epsilon \to 0$, $\log \epsilon \to -\infty$, so as ϵ is reduced, the influence of the zeroes on the analysis might become substantial. Reducing ϵ can also bring geometric means arbitrarily close to zero, which may be a further unappealing feature of this approach.

An alternative approach is to enquire if the zeroes really are zero. In many datasets of otherwise positive, skewed values, zero will mean that the observation is below some level of detection. If the level of detection can be identified, then an approach that takes this into account may be more satisfying and satisfactory. How such information is taken into account is likely to depend on the level of sophistication that the analyst believes is required. It might range from adding a value of half the limit of detection to the zeroes (which might be thought of as the expected value if the true observation were uniformly distributed between the limit of detection and 0), to some form of censored regression, treating the zeroes as censored values that are less than the limit of detection,

A final, more statistical, method is to add a parameter to each observation and estimate it from the data. Box & Cox (Box and Cox, 1964) did include a shifted version of their transformation, namely

$$y^{(\lambda_1,\lambda_2)} = \frac{(y+\lambda_2)^{\lambda_1} - 1}{\lambda_1},$$

with $y^{(0,\lambda_2)} = \log(y+\lambda_2)$. However, it is noteworthy that the authors never applied this shifted form of their transformation to any of the illustrations in their paper. The approach can be helpful, but estimation of such shift parameters can be challenging, especially for likelihood inference (Atkinson, 1985, p. 185).

References

A C Atkinson. *Plots, Transformations, and Regression*. Oxford University Press, Oxford, 1985.

G E P Box and D R Cox. An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, series B*, 26:211–252, 1964.

Kai Lai Chung. *A Course in Probability Theory*. Academic Press, New York, 1974.

Peter J Huber. *Robust statistics*. Wiley series in probability and mathematical statistics. Wiley, New York, 1981. ISBN 0471418056.

Edmond S-W Ng, Karla Diaz-Ordaz, Richard Grieve, Richard M Nixon, Simon G Thompson, and James R Carpenter. Multilevel models for cost-effectiveness analyses that use cluster randomised trial data: An approach to model choice. *Statistical Methods in Medical Research*, 25(5):2036–2052, 2016. doi: 10.1177/0962280213511719.

Suzanne Spence, Jennifer Delve, Elaine Stamp, John N. S. Matthews, Martin White, and Ashley J. Adamson. The impact of food and nutrient-based standards on primary school children's lunch and total dietary intake: A natural experimental evaluation of government policy in england. *PLOS ONE*, 8(10), 2013.

Li Su, Brian D. M. Tom, and Vernon T. Farewell. Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics*, 10(2):374–389, 2009. doi: 10.1093/biostatistics/kxn044.

Simon G. Thompson and Richard M. Nixon. How sensitive are cost-effectiveness analyses to choice of parametric distributions? *Medical Decision Making*, 25(4):416–423, 2005. doi: 10.1177/0272989X05276862.

Appendix

The temptation to give the following rather eccentric proof that $G_n \leq A_n$ is irresistible.

Let A_n, G_n be, respectively, the arithmetic and geometric means of the n positive numbers $x_1, \ldots, x_n$, and let $\mathscr{P}(n)$ be the proposition that $G_n \leq A_n$, with equality iff all the x_i are equal.

Now, $\mathscr{P}(1)$ is trivially true and $\mathscr{P}(2)$ follows from noting that $(\sqrt{x_1} - \sqrt{x_2})^2 \geq 0$.

The usual inductive step, $\mathscr{P}(n) \Rightarrow \mathscr{P}(n+1)$ is surprisingly awkward. However $\mathscr{P}(n) \Rightarrow \mathscr{P}(n-1)$ is straightforward: supplement the values $x_1, \ldots, x_{n-1}$ with $\bar{x}$, the arithmetic mean of the $n-1$ values. Applying $\mathscr{P}(n)$ to $x_1, \ldots, x_{n-1}, \bar{x}$ gives an expression that readily simplifies to $G_{n-1} \leq A_{n-1}$, i.e. $\mathscr{P}(n-1)$.

This isn't much use for a proof by induction, so a second step is needed. This is $\mathscr{P}(2^n) \Rightarrow \mathscr{P}(2^{n+1})$.

Consider the set of numbers $S = \{x_1, \ldots, x_{2^{n+1}}\}$ and, by a slight abuse of notation, let A_1 be the arithmetic mean of the first 2^n of these numbers and A_2 the same quantity for the second 2^n numbers. Let G_1, G_2 be the corresponding geometric means. As we know that $\mathscr{P}(2)$ is true

$$\tfrac{1}{2}(A_1 + A_2) \geq \sqrt{A_1 A_2} \geq \sqrt{G_1 G_2}$$

where the final inequality arises from $G_j \leq A_j$ because we assume $\mathscr{P}(2^n)$. The left hand expression above is the arithmetic mean of all values in S, whereas the right hand expression is their geometric mean, so the above shows $\mathscr{P}(2^{n+1})$ is true.

This last proof shows that $G_n \leq A_n$ for an infinity of n, and $\mathscr{P}(n) \Rightarrow \mathscr{P}(n-1)$ allows the gaps to be filled in!