Appendix A

The Lognormal distribution

A.1 Introduction

The lognormal distribution has been proposed as a model for shot length (SL) distributions by Barry Salt, the chapter *The Numbers Speak* in Salt (2006) being the fullest exposition (see, also, Salt, 2011). There has been some debate about how generally applicable this model is (Redfern, 2012a) and there are passing references to this in the text. My current view on this is that Redfern's criteria for accepting an SL distribution as lognormal is too stringent, and that to accept Salt's view means accepting an idea of what constitutes an acceptable approximation to lognormality, where what is meant by 'approximation' is not very precisely defined.

The idea is important and useful. If lognormality does apply it helps establish a 'norm' against which deviancy that merits exploration can be identified. This, and the concision of description offered, is sufficiently attractive that I think it worth retaining the idea if possible and erring on the side of interpreting 'approximation' in a generous way.

Even if not generally applicable, the lognormal model forms a useful ideal that might be exploited as a basis for pattern recognition, classifying SL distributions according to the different kind of departures from lognormality that they exhibit. This is most easily done after transforming SLs logarithmically when, in the ideal case, they then have a normal distribution. Most people probably find it easier to make judgments about departures from normality than lognormality.

I have noticed, using the sample of 150 (mostly) Hollywood films from 1935-2005 put together by Cutting *et al.* (2010), that the most obvious departures from normality after log-transformation are usually for the earlier films, where the ASLs are longer. Later films tend to be much less obviously non-normal after transformation; where they are it is usually evidenced by slight right-skewness and aberrant tail behaviour. Elsewhere (Baxter, 2012a) I've characterised 'obvious' non-normality after transformation as 'lumpiness', a rather loose concept but embracing distributions with more than one mode at the obvious end. Where distributions don't look normal but aren't 'lumpy' either, it is usually because after log-transformation they exhibit right-skewness of varying degrees of severity that the transformation is supposed to have removed.

I suspect, though more detailed investigation is needed, that it would be possible to develop a classification of SL distributions into 'types' that goes beyond a lognormal/not-lognormal dichotomy. In any event, and for whatever reason, the ideas of lognormality and normality after log-transformation are important, and referenced at lots of points in the text. What follows collects together the mathematical details concerning both distributions, interpretation of parameters in them and their estimation. Some attention is paid to notational distinctions, not always made in the literature, which can lead to confusion if neglected. An attempt is made to explain what logarithmic transformation does for your data; the idea is not what many people would regard as 'natural' but it is a useful one that anyone seriously interested in cinemetric data analysis should try to become comfortable with.

A.2 The lognormal and normal distributions

A.2.1 Definitions

The mathematical form of the lognormal probability density function is

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2\sigma^2}(\ln x - \mu)^2\right] \tag{A.1}$$

where x is strictly positive. The distribution is skew, bounded below by zero and tails off to a density of zero to the right as x approaches infinity. It is completely defined by the two *parameters* μ and σ.

Other ways of writing (parameterizing) equation (A.1) can be used such as

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2\sigma^2}(\ln x/\Omega)^2\right] \tag{A.2}$$

where $\ln \Omega = \mu$. This is the form used by Salt (2006: p.391).

Let X be a random variable with a lognormal distribution and write

$$X \sim \mathrm{LN}(\mu, \sigma^2)$$

to symbolize this. The logarithm of a lognormal variable, $Y = \ln X$, has a normal distribution and we write

$$Y \sim \mathrm{N}(\mu, \sigma^2)$$

where the probability density function of Y is

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2\sigma^2}(y - \mu)^2\right] \tag{A.3}$$

which is also completely specified by μ and σ and is singly-peaked and symmetric about μ.

A.2.2 Parameter interpretation

It is simplest to start with the normal distribution, for which the *mean* is μ and *standard deviation* σ. The *variance*, σ^2, is the square of the standard deviation. The *median* and *mode* of the normal distribution are the same as the mean, μ, but this is not generally true for non-symmetric distributions.

The mode, median, and mean are *measures of location*, the mode being the point at which the maximum value of the density occurs, the median being the point that has 50% of the distribution either side of it, and the mean being the centre of gravity of the data. The median can be referred to, though this is less common, as the 50th *percentile*. In common usage[1] the mode is sometimes called the most 'popular' or 'common' value, and the median the 'middle' value. The mean presents more difficulty since 'centre of gravity' is not widely understood; the undefined term 'average' is often substituted and often used sloppily without any clear indication of what is intended, or where it is confused with the median. The cinemetrics literature is not immune to this.

The standard deviation is a *measure of dispersion* with the useful interpretation, for the normal distribution, that just over two-thirds of the distribution lie within one standard deviation of the mean, and about 95% lies within two standard deviations. Again, this is not generally true.

The mean and standard deviation are often used in conjunction to summarise the location and dispersion of a set of data, whether normal or not. The median is the 50th percentile of the data; percentiles for other values can be defined; for example the 25th percentile has 25% of the distribution to its left and 75% to the right, and is sometimes called the first *quartile*. The

[1] That is, students I have taught; journalists (tabloid and broadsheet); etc.

75% percentile, the third quartile, can be similarly defined. The difference between them, the interquartile range (IQR), is often used as a measure of dispersion in conjunction with the median as a measure of location.

The *coefficient of variation* is a standardised (scale-free) measure of dispersion, defined as the standard deviation divided by the mean, so that for the normal distribution

$$\mathrm{CV} = \sigma/\mu.$$

For the lognormal distribution there is not a simple exact equality between population parameters and readily understood characteristics of interest descriptive of the population. Nor is there a simple equality between the mean, median and mode, which are

$$\mu_L = \exp(\mu + \sigma^2/2) = \Omega \exp(\sigma^2/2)$$

$$\Omega = \exp(\mu)$$

$$\sigma_L^2 = \left[exp(\sigma^2) - 1\right] \exp(2\mu + \sigma^2).$$

The mean, μ_L, is the 'average shot length', what is often denoted by ASL being its estimate. Note that if lognormality hold exactly there is a oerfect linear relationship between the mean and median. The median, Ω, is also the *geometric mean* of the data, to which we return in the next section. The variance of the lognormal distribution and CV are

$$\sigma_L^2 = \left[exp(\sigma^2) - 1\right] \exp(2\mu + \sigma^2).$$

$$\mathrm{CV}_L = \sqrt{exp(\sigma^2) - 1}$$

Some of the confusion that can arise in working with the lognormal disribution stems from the fact that, unlike the normal, the fundamental parameters that define it mathematically, μ and σ, don't equate simply with aspects of the distribution (or, in real life, the data) that one might think one has a 'feel' for. The median and CV are functions, respectively of μ and σ only.

Unlike the normal distribution the parameter σ, does not equate with dispersion in any simple way. In the context of the lognormal it is often referred to as a *shape* parameter, but the shape of the lognormal, as I suspect many would comprehend it, depends on both σ and μ. The lognormal can be described mathematically in terms of two numbers, μ and σ, which don't have the 'direct' interpretation that they have for the normal.

It can be emphasised that even in the ideal case of the lognormal *two* quantities are needed to describe the distribution fully. They should involve both μ and σ and this pair (μ, σ) is one obvious possibility. Baxter (2012a) suggests (μ_L, Ω); Salt (2011) suggests (μ_L, σ_L). There is no mathematical basis for preferring any of these, which in practice must be estimated, so whatever is most convenient to calculate, or most usefully interpretable in context, are reasonable grounds for choice, if one has to be made.

A.2.3 Models and estimates

Models

The lognormal and normal distributions are idealised *models* for data, of value practically because many different kind of real-life data *approximate* the distributions sufficiently well that the models, and their properties, can be used as the basis for very concise descriptions of the data aiding, in some circumstances, an understanding of the processes that generate the data.

Models are (almost) *never* correct. This is obvious for the theoretical normal and lognormal distribution, which disappear into infinity in both directions for the normal and to the right for the lognormal. Real data can't do this[2]. In practice the most one can hope for is that over the relevant

[2]Taking SLs as an example and Googling 'longest takes' brings up frequent references to Hitchcock's *Rope* (1948) with a maximum SL that, at just over 10 minutes, is neither a contender nor anywhere near infinity.

range of a distribution observed data 'approximately' match that which would be predicted by the model under consideration.

I suspect differences in what is to be understood by 'approximately' lie at the heart of disputes about whether SL data do tend to have lognormal distributions (Salt, 2006[3], 2011; Redfern 2012a) but this isn't pursued here. It's possible to conceive of any analysis of SL data as occupying one or more levels of a hierarchy, of which model-fitting is the highest. At the lowest level, and possibly most important, is graphical analysis.

At an intermediate level the computation of descriptive statistics – the ASL, median , standard deviation, IQR, whatever – are attempts to usefully quantify aspects of the data that provide insights into it, one hopes, from what might be called a 'filmic' perspective. Such efforts can be conceived of as either of merit in their own right or, additionally, as estimating characteristics of a distribution of which the SL data are a sample[4]. In this latter case it is not necessary to assume any particular distribution that underlies the data; if one does make such an assumption and wants to check it the highest level of the hierarchary, model-fitting, is reached.

Estimation and notation

Differentiating between characteristics *estimated* from a set of data and the analogous characteristics or parameters of the hypothesised and idealised underlying model is important. The latter are unknown and estimates, which are calculated and known quantities, are typically meant to be usefully informative about what is unknown. It is important to maintain a notational distinction between the two, which can be confusing when done properly, but is even more so when not. As an example take the average shot length (ASL) which can be written variously, as

$$\text{ASL} = \hat{\mu}_L = \bar{x} = \frac{\sum x}{n}$$

where the last term just means to add the individual shot lengths (the x) and divide by the number of shots, n. The circumflex, or 'hat' notation is applied to distinguish estimated sample quantities from their unknown population counterparts. Conventionally, the latter are indicated using letters from the Greek alphabet. Thus $\hat{\mu}_L$ estimates the unknown population mean μ_L, the subscript, L, distinguishing this from the population parameter μ.

Other notation is in common use and $\bar{x}$ is standard notation for the mean of a set of numbers denoted by x. Terminology, alas, can also be confusing. Salt, in coining ASL, was well aware that 'arithmetic mean' was the strictly correct terminology but opted for 'average' on the grounds that it was what his intended audience could live with (Salt, 2012)[5]. The qualifying adjective, 'arithmetic' is often dispensed with but it can be useful to retain, as other kinds of mean can be defined mathematically, and appear in cinemetric studies.

If similar calculations are applied to *logarithmically transformed* data

$$\hat{\mu} = \bar{y} = \frac{\sum y}{n}$$

which provides a simple and direct way of estimating μ, which may be used for fitting the lognormal model. Note that $y = \ln x$ is used in the calculations. The parameter σ, which happens to be the

Alexander Sokurov's *Russian Ark* (2002) with a single take of 96 minutes is more impressive but, as these things go, nowhere near infinity. There is a serious point here. Arguments against the use of the ASL as a measure of 'film style' sometimes focus on the fact that it is not a robust statistic, meaning in theory that it has an asymptotic breakdown point of zero, meaning that you can make the ASL as large as you like by adding an arbitrarily large and hypothetical SL to the data. So add an SL approaching infinity and the ASL does so to. The relevance of this theoretical argument, when not only do real SLs not approach infinity, but struggle to get over, say, four minutes, might be questioned.

[3]In the chapter *The Numbers Speak* in the book.

[4]I'm aware of objections (Salt, 2012) to treating SL data as samples rather than the entire population, but won't pursue the issues this raises here.

[5]That confusion does arise is evidenced by publications in the journal literature, where SLs are analysed, that use ASL and 'median' as if they are the same thing (see Baxter, 2012a, for references).

standard deviation of the normal distribution, is most simply estimated from

$$\hat{\sigma}^2 = s^2 = \frac{\sum(y - \bar{y})^2}{n - 1}$$

where s as an estimate of standard deviation is also in common use[6]. Replacing y with x in the above provides an estimate of σ_L^2.

The median of the lognormal, Ω, can be estimated in various ways. Thinking of this as a population characteristic, the obvious estimate of it is the sample median, $\hat{\Omega}_M$, where the subscript M is introduced to distinguish it from other possibilities. One such, as noted in passing above, is to use the *geometric* mean of the data, that is $\exp(\hat{\mu})$ where $\exp()$ is the exponential function and $\hat{\mu}$ is the arithmetic mean of the logged data. The geometric mean of a set of n numbers is defined as the nth root of the product of the numbers so

$$\hat{\Omega}_G = (\prod x)^{(1/n)}$$

giving

$$\ln \hat{\Omega}_G = \ln [(\prod x)^{(1/n)}] = \frac{\sum \ln x}{n} = \frac{\sum y}{n} = \hat{\mu}$$

and back-transforming to

$$\hat{\Omega}_G = \exp(\hat{\mu}).$$

The mathematics here is for the record and can be ignored, so long as it's appreciated that more than one estimate of the median is available. Parkin and Robinson (1993) study four possibilities, including the two mentioned here which are the simplest. Their conclusions are not simply summarised and neither of the two estimators considered here are uneqivocally better than the other.

The comparative use of the statistics is interesting. They should give similar results for large samples *if* the underlying distribution is lognormal. Redfern's (2012a) study of 134 films suggests shows a reasonably systematic difference in results for the two estimates which casts doubt on the view that the lognormal distribution is generally applicable and shows that the issue of choice is not an 'academic' one[7].

A.3 Why logarithms?

The use of logarithms is unavoidable in discussing the lognormal distribution. I suspect, with the long-ago introduction of pocket calculators, logarithms are no longer taught in schools as they once were, and it wasn't necessarily that easy then.

What is involved is a transformation from one set of numbers to another, $X \rightarrow Y$ say, or

$$y = f(x) = \ln x$$

. Back transformation, $x = \exp(y$ is possible so you can always get from one set of numbers to another. The transformation is *monotonic* meaning that it retains the same ordering of numbers.

In terms of application to SLs a log-transformation does several things. SLs are bounded below by zero and distributions are skewed, often with a long tail. For some statistical purposes the lower bound of zero is an inconvenience and the log-transformation frees the data from this constraint so that negative numbers are possible and the data are (theoretically) unbounded. The values of SLs, measured in seconds, differ by orders of magnitude, meaning that a shot of 1 second is ten times longer than a 1 deci-second (0.1 second)shot; one of 10 second is 10 times greater than a 1

[6] In some statistics texts the divisor n rather than $(n - 1)$ is used. The reasons for this, where it is knowingly and correctly done, are not of concern here, since n is usually so large the difference has no effect on calculations. Similar comments apply when, as sometimes happens, $\hat{\sigma}^2$ and s^2 are defined with different divisors.

[7] Even if not strictly true I think the question of whether or not lognormality is adequate as an *approximation* in many cases remains open.

second shot and so on. This can be inconvenient for both statistical analysis and graphical display because results can be dominated, not necessarily beneficially, by the larger SLs.

The effect is removed by logarithms which transform SLs to the same order of magnitude so that (using logarithms to base 10) the logarithms of 0.1, 1, 10, 100, become -1, 0, 1, 2. (Logarithms to other bases can be used and the numbers will differ, but the difference between them will be constant[8].) What this means, in graphical terms, is that equal weight is given to short and long SLs so that, for example, SL data that are lognormally distributed are symmetrically and normally distributed after log-transformation. Departures from log-normality on the original scale will appear as departures from normality on the log-scale, and my guess is that most people find it easier to make visual judgments about such departures with a normal distribution as the reference.

For making such assessments, including formal tests of log-normality, log-transformation is not only a possibility, but also the 'natural' thing to do. I'd also contend that it often more useful to compare SL distributions on a log-scale, and several examples are provided in the text.

Log-transformation also has the effect of downweighting the influence of outliers if they exist, without sacrificing information about the fact that some SLs are noticeably bigger than others – information that is 'lost' if rank-transformations are used, for example. In this sense using a log-transformation has some of the advantages claimed for 'robust' methodologies, without the sacrifice of information inherent in the latter usage.

The price paid for the often considerable advantages of using log-transformation are offset by the discomfort induced by the abstraction involved – a distancing from the 'reality' of the raw data. A shot, as I understand it, is composed of a continuous sequence of frames, broken at either end by a cut or some other recognisable form of transition. If you're looking for it you can recognise a shot when you see one and get some 'feel' for its length; you don't think or 'feel' in terms of the logarithms of SLs.

What's often sensible is to take a deep breath, do an analysis using logarithms then, if this causes discomfort, translate back to the original scale and see if the conclusions drawn from the log-based analysis convince you.

A.4 Other normalising transformations

A logarithmic transformation of the data is often useful for getting a better 'look' at some aspects of the data. Beyond this there is a specific interest in using it to investigate the claim that a majority of films have SL distributions that approximate lognormality. This involves transforming the original data, X to $Y = \ln X$ and investigating normality of Y.

Where this is not successful it is worth asking whether the data can, nevertheless, be transformed to normality using other than a logarithmic transformation. At the cost of what is only a slight additional mathematical complexity, this is readily explored. A rather simplistic way of thinking about the lognormal transformation is that it 'squeezes' in the long right tail, allowing more freedom to the left tail, in an attempt to get the data to 'behave' normally.

It sometimes works, but not always – the data may be recalcitrant, but can look as a bit of an extra squeeze might remedy matters. The thought could also occur that you might have tried squeezing a bit harder in the first place. This is, more-or-less, what's attempted with the two transformations to be described.

The *Box-Cox* (BC) transformation applies differential 'force' from the word go; mathematically the amount of force needed to get the data to behave as well as possible (it may still end up behaving badly) is determined by an extra parameter, λ, that is where the additional complication comes in. This value is close to zero if the data are acceptably lognormal, and tends to become more negative as more force needs to be exercise to achieve normality (this observation is based on SL data I've looked at and does not apply generally to other kinds of data).

[8] Logarithms can be to different bases; natural logarithms, to base e, have been used elsewhere in the text. These are sometimes written as $\log_e$ in contrast to $\log_{10}$ for base 10 logarithms. The convention used in the notes is $\ln()$ for the former and $\log()$ for the latter; the convention is widespread but not universal.

The other approach investigated initially applies equal force to the data, via a gentle log-transformation, but then applies an additional and differential squeeze to recalcitrant data sets. What motivates this idea is that initial log-transformation is intended to remove skewness, but often leaves the data still lookning skew. A second log-transformation might seem in order but mathematical complications arises with this. To get round the problems here the use of the *Yeo-Johnson* (YJ) transformation was explored. It looks (and is) more complicated than the BC transformation, but depends on only a single extra parameter.

The BC and YJ transformations are outlined mathematical below. Their application, particularly of the latter, is explored in the text. What's involved is a double-transformation or double-'squeeze' of the data, with uniform squeezing at the first stage and differential squeezing at the second stage.

What can potentially be achieved is (a) a categorization of SL distributions into those which can and cannot be transformed to normality and (b) for those which can, a characterisation of the amount of 'force' needed to achieve normality. The substantive interpretation of this is discussed in the relevant chapters of the text.

A.4.1 Box-Cox (BC) transformations

Power transformations of the form X^λ have been used in statistical analysis for some time; the slightly more 'complex' version studied by Boz and Cox (1964) (see Sakia, 1992, for a review of early work)

$$X^\lambda = \frac{X^\lambda - 1}{\lambda}$$

is that used here. In applications a transformation, λ, is sought that makes the data as nearly normal as possible. It is defined for $X > 0$, in such a way as to ensure continuity, with a value of $\lambda = 0$ equivalent to a log-transformation. That is, the log-transformation that can be used to investigate the lognormality of SL distributions is a special case. If it is concluded that, for any particular film, lognormality does not hold it is possible to see if some other value will induce normality.

Theory and advice on practical application is readily enough found in the literature; for practical analysis a variety of functions in `R` can be found and the `powerTransform` and `bcPower` functions from the `car` library are those used in the text.

A.4.2 Yeo-Johnson (YJ) transformations

The Yeo-Johnson transformation looks somewhat nastier, though at heart it is the same kind of thing as the BC transformation, but modified to deal with negative values. Here it is applied after an initial log-transformation, $Y = \ln X$. This may induce negative values, which is why the extra complexity of the YJ transformation is needed. It is defined as

$$\begin{array}{ll} (Y^\lambda + 1) - 1]/\lambda & \text{if } \lambda \neq 0; y \geq 0 \\ \ln(Y+1) & \text{if } \lambda = 0; y \geq 0 \\ -[(Y+1)^{2-\lambda} - 1)]/(2-\lambda) & \text{if } \lambda \neq 2; y < 0 \\ -\ln(-Y+1) & \text{if } \lambda = 2; y < 0 \end{array}$$

and can be implemented using the `powerTransform` function from the `car` package in `R`. The meaning of λ can be hard to interpret; the transformation can be viewed as a generalisation of the BC transformation that amounts to applying different power transformations to negative and positive Y modified($|Y| + 1$) or $(Y + 1)$, with powers $(2 - \lambda)$ or λ, the additional complexities ensuring continuity.