

# The lognormal distribution is not an appropriate parametric model for shot length distributions of Hollywood films

Nick Redfern

## Abstract

We examine the assertion that the two-parameter lognormal distribution is an appropriate parametric model for the shot length distributions of Hollywood films. A review of the claims made in favour of assuming lognormality for shot length distributions finds them to be lacking in methodological detail and statistical rigour. We find there is no supporting evidence to justify the assumption of lognormality in general for shot length distributions. In order to test this assumption we examined a total of 134 Hollywood films from 1935 to 2005, inclusive, to determine goodness-of-fit of a normal distribution to log-transformed shot lengths of these films using four separate measures: the ratio of the geometric mean to the median; the ratio of the shape factor  $\sigma$  to the estimator  $\sigma^* = \sqrt{2 \times \ln(\bar{x}/M)}$ ; the Shapiro-Francia test; and the Jarque-Bera test. Normal probability plots were also used for visual inspection of the data. The results show that, while a small number of films are well modelled by a lognormal distribution, this is not the case for the overwhelming majority of films tested (125 out of 134). Therefore, we conclude there is no justification for claiming the lognormal distribution is an adequate parametric model of shot length data for Hollywood films, and recommend the use of robust statistics that do not require underlying parametric models for the analysis of film style.

**Keywords:** lognormal distribution, goodness-of-fit, film style, shot length distribution

*Je n'avais pas besoin de cette hypothèse-là*

Pierre-Simon Laplace

## 1. Introduction

Although the most frequently cited statistic of film style is the average (mean) shot length (ASL), the distribution of shot lengths in a motion picture is definitely not normal. The distribution of shot lengths in a motion picture is typically positively skewed and contains a number of outlying data points whilst being bounded by zero at the lower end. Given the skewed nature of this data, a logarithmic transformation may be appropriate before proceeding with statistical analysis in order to 'normalise' the distribution by removing the skew and the influence of outliers (Quinn & Keogh: 64-67). The data can then be summarised by finding the parameters for the underlying lognormal distribution, and can be analysed using parametric methods that assume data is normally distributed with the results transformed back to the original scale. The use of the lognormal distribution for describing the shot lengths in a motion picture in general has been suggested in two papers: one by Barry Salt (with an additional commentary on line), and one by Jordan De Long, Kaitlin L. Brunick, and James E. Cutting.

In this paper we review the claim that a lognormal distribution is an appropriate parametric model for the distribution of shot lengths in a motion picture. In the next section we introduce the lognormal distribution and note some relevant features to this study. In section three we review the claims that this distribution is an appropriate model for shot lengths in Hollywood films, focussing on the methodology employed and the conclusions derived. In section four we test these claims against the shot length distributions of a sample of 134 films using a range of different methods to determine if lognormality is an appropriate assumption. From these results we draw some conclusions about the appropriate use of statistics in the analysis of film style.

## 2. The lognormal distribution

The distribution of shot lengths in a motion picture is asymmetric and exhibits positive skewness. In part this is due to a number of shots that are much longer than the majority of others. An additional reason is that the distribution is bounded on one side by zero: no shot can be equal to or less than 0 seconds long. This means the distribution cannot be approximated by a normal distribution since this would imply that some shots have negative length, and this is obviously impossible. Therefore the mean shot length and the standard deviation cannot be used to give an accurate description of film style since these are the parameters of the normal distribution.<sup>1</sup> In these circumstances we may suspect that a skewed probability distribution will prove to be an adequate model for the data, and given the lower limit of the distribution of the data is bounded by zero that the lognormal distribution will fulfil this role. In applying a logarithmic transformation we do so in the expectation that the resulting data set will be symmetrical and that the influence of outliers will be removed (i.e. the transformed variable will be normally distributed), and that the dependence of the standard deviation on the mean will be eliminated.

The lognormal distribution is defined in relation to the normal distribution: a random variable  $X$  is lognormally distributed if its logarithm  $Y = \log(X)$  is normally distributed. Therefore, the lognormal distribution is a continuous probability distribution with the density function

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(\log x - \mu)^2\right], \quad 0 < x < \infty,$$

where  $\mu$  is the arithmetic mean of  $\log(X)$  and  $\sigma$  is the standard deviation of  $\log(X)$ . This is true irrespective of the base of the logarithm, but throughout this paper we will use the natural logarithm (i.e. the logarithm to base  $e$ ). Kleiber and Kotz (2003: 107-145), Burmaster and Hull (1997), and Limpert, Stahel, and Abbt (2001) provide detailed reviews of the properties of the lognormal distribution.

Back-transforming  $\mu$  into the original scale of the data gives the *geometric mean* ( $G$ ), which is approximately equal to the *median* ( $M$ ) if the data are lognormal since the median of a lognormal distribution is at  $x = \exp(\mu)$ . In this context  $\exp(\mu)$  is a location factor. As the distribution is right-skewed both the geometric mean and the median are less than the

---

<sup>1</sup> For example, based on data from the sample discussed below, the mean shot length of *The Apartment* is 16.1 seconds with a standard deviation of 19.1 seconds, implying that 20% of the shots in this film are less than or equal to 0 seconds.

arithmetic mean ( $\bar{x}$ ) of the data in the original scale. However, if a random variable is lognormally distributed then we can estimate the value of the arithmetic mean by exploiting a mathematical relationship between the arithmetic mean, the median, and the shape factor of the lognormal distribution, where  $\bar{x} = M \times \exp(0.5\sigma^2)$ .

Since  $X$  is lognormally distributed if  $Y = \log(X)$  is normally distributed, it is easy to analyse the distribution of data as all we need do is apply a logarithmic transform and then use statistical methods based on normal distributions. Testing the goodness-of-fit of a lognormal distribution to data is therefore equivalent to testing the goodness-of-fit of a normal distribution to  $Y = \log(X)$ . The principles and methods of testing the goodness-of-fit of a normal distribution are well understood (see Thode 2002), and normality tests are widely available in statistical software packages.

### 3. The lognormal distribution and film style

In this section we examine the methodology and conclusions behind the claims of Salt and of De Long, Brunick, and Cutting that a lognormal distribution adequately models the frequency distribution of shot lengths in a motion picture.

#### 3.1 Barry Salt, 'Let the numbers speak' (2006) and 'The metrics in Cinematics' (2011)

The strongest claim for the use of lognormal distribution in analysing film style has been put forward by Barry Salt (2006: 389-396; 2011), who asserts the generality of the lognormal distribution for the shot length distributions of all films. On the basis of this assertion he makes a series of subsequent general claims about shot length distributions: that the ASL is an informative statistic even with heavily skewed shot length distributions because  $\bar{x} = M \times \exp(0.5\sigma^2)$  under lognormality; that the characteristic shape factor of shot length distributions is  $\sigma = 0.9$ ; and that the ASL and  $\sigma$  are 'fairly independent' up to  $ASL \approx 20$  seconds.

There are three principal problems with Salt's underlying claim of lognormality. The first problem is the sample of films on which these claims are based. In Salt (2006) the sample used contains a total of just 18 films, and one of these uses data from only the first 40 minutes of the film. Of the films in the sample two are silent films (1 from 1916 and 1 from 1924), 8 were released in the 1930s, four from the 1960s, and one each from the 1940s, 1950s, 1980s, and 1990s. Six of the films are French, one is German, one is Brazilian, and the remainder are American. This sample cannot be considered to represent any population of films sorted by historical era or by nation, and it is unclear on what basis the results are generalized to other films. Salt (2011) provides further individual examples but again there is no attempt to systematically analyse a large sample of films that represents a defined population.

The second problem is the method by which Salt measures goodness-of-fit. Salt's measures goodness-of-fit by plotting the frequency distribution of shot lengths in a histogram with the predicted frequency distribution from a lognormal distribution on logarithmic plotting paper. Goodness-of-fit is then determined using the coefficient of determination ( $R^2$ ), with values close to 1 considered as evidence the data is well fitted by a lognormal distribution. This is described as the 'standard method' of determining the relationship between two variables, and the squared correlation between the observed and theoretical quantiles of a distribution is a powerful method of determining goodness of fit when combined with normal probability plots (see sections 4.1.3 and 4.1.4). The use of  $R^2$  here, however, is incorrect. Sorting the data changes its structure so that data values are no longer

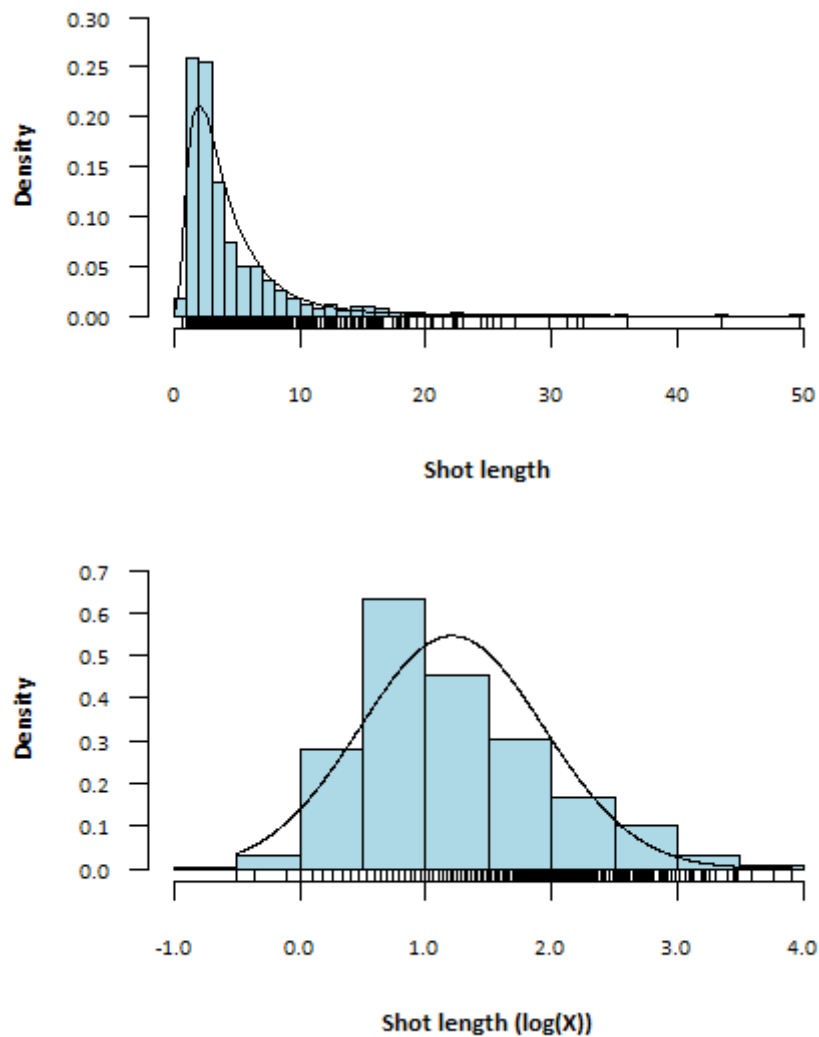
independently and identically distributed, and it is necessary to take into account the high correlation and heteroscedasticity of the order statistics. In these circumstances ordinary least squares regression is not appropriate since the relationship between the ordered data and the ordered theoretical values is always monotonically increasing, and it is necessary to interpret the results in the context of a generalized least squares model (Cullen & Frey 1999: 135). There is no indication given by Salt this is the case, and since these features may result in high values of  $R^2$  even if the underlying distribution is not lognormal the results presented are unreliable. Additionally, since the data is binned only a limited amount of information is used and the degree of correlation between the theoretical and observed values is dependent on the size of the bins chosen for the histogram rather than the data itself.<sup>2</sup> Finally, no decision rule or null distribution is given for the  $R^2$  statistics and so what constitutes goodness-of-fit in this context is not defined.

Problem number three arises in the use of graphics to support the claim of lognormality. Goodness-of-fit in both articles is represented visually using histograms with fitted lognormal distributions on an arithmetic scale rather than as histograms of the transformed data with a fitted normal distribution on a logarithmic scale. This gives a misleading impression of the goodness-of-fit and Figure 1 illustrates the difference this can make. The histogram of Salt's data for *Little Annie Rooney* (without titles) on an arithmetic scale clearly shows the positive skew and outliers of this distribution and the fitted distribution appears to be a reasonable fit. However, when the same data is plotted on a logarithmic scale the same data does not show the familiar bell-shaped curve of the normal distribution and is obviously still skewed after the data transformation has been applied. The normal distribution fitted to  $\log(X)$  is obviously a very poor fit. The histogram on the arithmetic scale may lead us to infer this data is lognormally distributed and for this film Salt (2011) reports that  $R^2=0.964$ , but examination of the log-transformed data shows this conclusion to be wrong (Shapiro-Francia: 0.9556,  $p = <0.01$ ).

There are, then, serious methodological issues regarding Salt's conclusions, which are based on a small, unrepresentative sample and flawed techniques for determining goodness-of-fit. Based on the evidence in this paper, the claim cannot be considered proven and certainly does not justify Salt's interpretation that the lognormal distribution is an appropriate parametric model for such data in general for shot length distributions.

---

<sup>2</sup> In his online commentary, Salt refers to Aitchison and Brown's (1957) monograph on the lognormal distribution where the plotting of the grouped cumulative frequencies on lognormal probability paper is described but he does not apply Geary's test or the  $\chi^2$  test of normality to the log-transformed data even though these are discussed in the same text (see Aitchison & Brown 1957: 28-36). In the circumstances, the  $\chi^2$  test would seem a more obvious choice for determining goodness-of-fit than  $R^2$ . However,  $\chi^2$  is not recommended as a normality test since binning the data is inherently wasteful and affects the test statistic, and there are many other normality tests available that are considerably more powerful (D'Agostino 1986).



**Figure 1** Histograms of shot length data for *Little Annie Rooney* (without titles) on an arithmetic scale with fitted lognormal distribution  $\text{Log-}\mathcal{N} \sim (1.2078, 0.7304^2)$  (top) and on a logarithmic scale with fitted normal distribution  $\mathcal{N} \sim (1.2078, 0.7304^2)$  (bottom). Source: [http://www.cinematrics.lv/movie.php?movie\\_ID=6692](http://www.cinematrics.lv/movie.php?movie_ID=6692), accessed 27 January 2011.

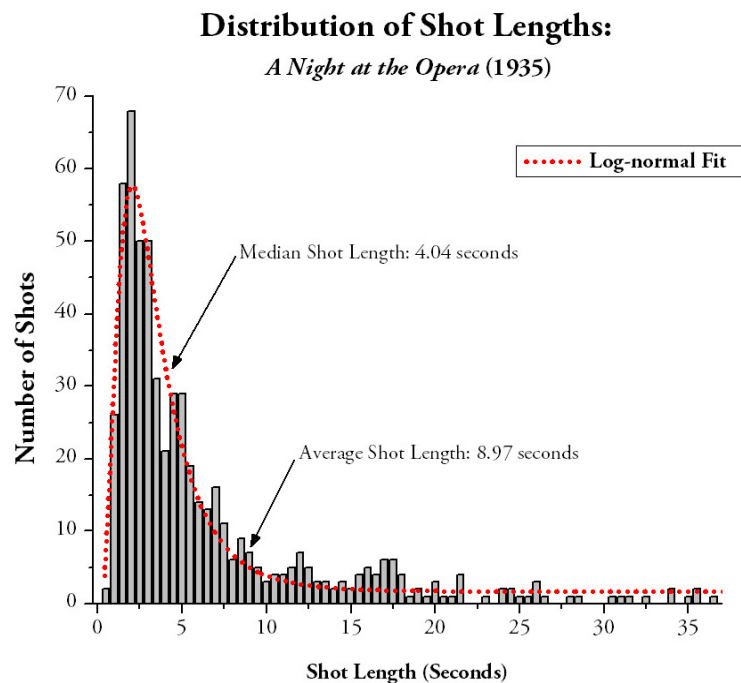
### 3.2 De Long, Brunick, and Cutting, 'Film through the human visual system: finding patterns and limits' (2012)

The claim that shot length distributions are lognormal has been repeated by De Long, Brunick, and Cutting (2012), whose research on films style is based on a sample of 150 high grossing films at the US box office sampled five years apart from 1935 to 2005, inclusive. They write

Despite being the popular metric, ASL may be inappropriate because the distribution of shot lengths isn't a normal bell curve, but rather a highly skewed, lognormal distribution. This means that while most shots are short, a small number of remarkably long shots inflate the mean. This means that the large majority of shots in a film are actually below average, leading to systematic over-estimation of individual

film's shot length. A better estimate is a film's Median Shot Length, a metric that ... provides a better estimate of shot length.

The appropriateness of assuming a lognormal distribution for shot length distributions is casually asserted but is not demonstrated. Although their sample included 150 films, they provide only one example of a shot length distribution with a fitted lognormal density functions to support this argument (see Figure 2). No other results are presented, and it appears the reader is expected to infer the generality of the lognormal distribution based on this single piece of evidence. Again, no goodness-of-fit tests are conducted to test the null hypothesis of lognormality for any data set. Like Salt, they use the histogram and density function of the untransformed data for *A Night at the Opera* to illustrate goodness-of-fit. The result is similarly misleading, and, as we shall see below, the distribution of shot lengths in *A Night at the Opera* is definitely *not* lognormal.



**Figure 2** Histogram of shot lengths in *A Night at the Opera* with a fitted lognormal distribution from De Long J, Brunick KL, and Cutting JE 2012 Film through the human visual system: finding patterns and limits, in JC Kaufman and DK Simonton (eds.) *The Social Science of Cinema*. New York: Oxford University Press: in press. This graph was downloaded from the online version of this paper available at <http://people.psych.cornell.edu/~jec7/pubs/socialsciencecinema.pdf>, accessed 18 January 2012.

If Salt's claims are the result of a methodologically flawed analysis, then De Long, Brunick, and Cutting do not appear to have used any methodology for assessing goodness-of-fit at all. They simply state that shot length distributions are lognormal, and present one chart as evidence that this is true in general. Naturally, we would not expect to find a paper presenting charts similar to Figure 2 for all 150 films in a sample, but we may reasonably expect some more detailed evidence in support of so general and unequivocal a statement as

that quoted above. Given this situation, we do not consider De Long, Brunick, and Cutting's paper to be any sort of evidence of the lognormality of the shot length distributions of motion pictures.

There is a clear difference between these two papers in their interpretation of the asserted lognormality of shot lengths and this also creates methodological problems for the statistical analysis of film style. Salt (2011) uses the lognormal distribution to justify retaining the mean shot length based on the relationship between the arithmetic mean, the median, and  $\sigma$  under lognormality, while De Long, Brunick, and Cutting adopt the opposite conclusion and refer to the lognormal distribution as justification for using the median shot length in place of the mean because it does not lead to the systematic overestimation of a film's cutting rate given a skewed data set. Therefore, although the claims made in these two papers may at first appear to be mutually supportive, they are in fact opposed in their fundamental approaches to analysing film style. Inevitably, the result is unnecessary confusion for the reader since we are expected to maintain two conflicting ideas based on the same reasoning for two statistics that lead to obscure and contradictory conclusions.

### **4. Is lognormality a reasonable assumption for shot length distributions of Hollywood films?**

Since the appropriateness of the lognormal distribution for shot lengths has not been demonstrated in this section we test this claim against a sample of Hollywood films to determine if it is in fact reasonable to make any such assumption.

#### **4.1 Methods**

We assessed the lognormality of the data using several different methods based on descriptive statistics, normal probability plots, and statistical tests of the null hypothesis of lognormality. Two different types of normality test were applied to the logarithms of the shot length data for each to avoid relying on a single test that may fail to identify some kinds of deviation from the assumed model. It is important to remember that failure to reject the null hypothesis is not proof that a lognormal distribution is an appropriate model. The decision on whether or not lognormality was assumed was based on consideration of all these factors together, and, if necessary, the use of additional graphical methods (e.g. histograms, density traces, etc).

All statistical analysis was conducted using **R** (version 2.13.0) and Microsoft Excel 2007.

##### *4.1.1 Sample*

The sample used in this study is that used by De Long, Brunick, and Cutting in their analyses of film style since this is the data on which they apparently base their claims regarding film style. The majority of these films are American, with a handful of them British, and cover seventy years of filmmaking and a wide range of genres and filmmakers. This data can be accessed via the Cinemetrics website: <http://www.cinemetrics.lv/index.php>. Although this sample comprises a total of 150 films we were unable to use all the films in our analysis due to rounding or what we assume are data entry errors. The minimum shot length for nine films was given as 0.0 seconds, while the minimum was less than 0.0s for seven films. These films were excluded as no conclusion could be reached regarding lognormality because logarithms exist only for real numbers strictly greater than zero. Therefore, the sample used in this study comprises a total of 134 films.

#### 4.1.2 Descriptive statistics

The two-parameter lognormal distribution is described by the parameters  $\mu$  and  $\sigma$ , where  $\log(X) \sim \mathcal{N}(\mu, \sigma^2)$ , and so we focus on estimates of these parameters as a means of assessing goodness-of-fit.

The cutting rate of a film is described in seconds, and we interpret the average shot length in arithmetic space even though transforming the data implies we conduct any analysis in logarithmic space. We are therefore interested in  $\exp(\mu)$  as a measure of location. Candidates for  $\exp(\mu)$  are the geometric mean ( $G$ ) and the median ( $M$ ). If the data are lognormally distributed then these statistics will be approximately equal, and  $\frac{G}{M} = 1$ . If this ratio differs from 1 we conclude the data is not lognormally distributed. As the median may be greater than the geometric mean any description of the ratios for the whole sample of films will underestimate the true average discrepancy and so we use the consistent ratio  $\frac{\max(G,M)}{\min(G,M)}$  to estimate the median discrepancy between these two estimates of  $\exp(\mu)$ .

Salt claims we should retain the ASL as a statistic of film style because its ratio to the median shot length allows us to derive the shape factor of a lognormal distribution that adequately describes the distribution of shot lengths in a motion picture. Therefore, we compare two estimates of the shape factor of the distributions:  $\sigma$  – the standard deviation of the log-transformed shot lengths (the maximum likelihood estimate); and the estimate derived from the ratio of the arithmetic mean ( $\bar{x}$ ) and the median ( $M$ ):  $\sigma^* = \sqrt{2 \times \ln(\bar{x}/M)}$ . If the data are lognormally distributed then the two shape factor estimates will be approximately equal and  $\frac{\sigma^*}{\sigma} = 1$ . If this ratio differs from 1 the lognormal distribution is not a good model for the data. Again, we use the consistent ratio  $\frac{\max(\sigma, \sigma^*)}{\min(\sigma, \sigma^*)}$  to determine the true size of the median discrepancy between estimates since  $\sigma$  may be greater than  $\sigma^*$ . Since Salt gives the shape factor on a logarithmic scale we follow this convention, and we do not use the multiplicative standard deviation ( $\exp(\sigma)$ ).

#### 4.1.3 Normal probability plots with log-transformed data

A normal probability plot is a scatter plot of the quantiles of the observed distribution (i.e. the log-transformed shot lengths of a film) against the expected quantiles of the theoretical distribution (the normal distribution). The points in the normal probability plot will show a strong linear pattern if the theoretical distribution is a good model for the observed quantiles. Fitting a reference line makes assessing the linearity of the plot easier, where the intercept and the slope of the line are the generalized least squares estimates of  $\mu$  and  $\sigma$ , respectively. Deviations from this line indicate the normal distribution is not an appropriate model for the data. See Burmaster and Hill (1997) for an overview of assessing lognormality with probability plots.



#### 4.1.4 The Shapiro-Francia test

The Shapiro-Francia (SF) test (1972) is a correlation-based goodness-of-fit test related to the normal probability plot. The test statistic is the squared correlation between the expected values of the normal order statistics ( $m_i$ ) and the observed quantiles ( $x_{(i)}$ ),

$$W' = \left[ \sum_{i=1}^n m_i x_{(i)} \right]^2 / \left[ \sum_{i=1}^n m_i^2 \times \sum_{i=1}^n (x_i - \bar{x})^2 \right]$$

$W'$  is equal to the square of the probability plot correlation coefficient and is therefore a measure of the linearity of a normal probability plot (Filliben 1975). Looney and Gullledge (1985: 76) point out the order statistics of the observed quantiles are highly correlated and heteroscedastic, and that the usual set of critical values used for interpreting a correlation coefficient do not apply in these circumstances.  $W'$  should not be confused with  $R^2$  since this may lead to flawed inferences arising from use of the wrong null distribution.

The Shapiro-Francia test was implemented using the **R** package `nortest` (version 1.0), which allows for samples of size  $5 \leq n \leq 5000$ . Lognormality was rejected at  $\alpha = 0.05$ .

#### 4.1.5 The Jarque-Bera Test

As sample sizes become very large it becomes increasingly likely that a statistical hypothesis will reject the null hypothesis in the presence of trivial deviations from the theoretical model. The number of shots in films from the sample range from 233 to 3077, and so we employ the Jarque-Bera test (1987) as an additional test of normality suitable for large sample sizes.

The Jarque-Bera test is a goodness-of-fit test based on the skewness and kurtosis of the normal distribution. Skewness describes the asymmetry of a distribution and kurtosis is a measure of the peak of the data relative to a normal distribution. For a normal distribution, skewness is equal to 0 (i.e. the distribution is symmetrical) and the kurtosis is equal to 3. The Jarque-Bera test compares the sample skewness and kurtosis to these values. The test statistic for a sample of size  $n$  is

$$JB = \frac{n}{6} \left( S^2 + \frac{1}{4} (K - 3)^2 \right),$$

where  $S$  is the sample skewness and  $K$  is the sample kurtosis. Under the null hypothesis, JB has an asymptotic  $\chi^2$  distribution with two degrees of freedom.

The Jarque-Bera test was implemented using the **R** package `tseries` (version 0.10-27). Lognormality was rejected at  $\alpha = 0.05$ .

## 4.2 Results

Table 1 presents the full set of results for each film. We find the lognormal distribution is not an appropriate model for 125 of the 134 films in the sample. Of the remaining nine films,

lognormality appears to be a reasonable assumption in eight cases. The results for one film (*Blood on the Sun*) are inconclusive and this film is discussed separately.

#### 4.2.1 Descriptive statistics

The first method employed was the comparisons of the ratios  $\frac{G}{M}$  and  $\frac{\sigma^*}{\sigma}$ . The ratios of the location estimates range from a minimum of 0.93 for *Harvey* to a maximum of 1.22 for *Annie Get Your Gun*. The median is greater than the geometric median in five cases, but is only substantially different for *Harvey*. The median of the consistent ratios of the location estimates is 1.08 (IQR: 0.07, 95% CI: 1.07, 1.09). The range of values for the ratio of the shape factor estimates is from a minimum of 0.89 for *Harvey* to a maximum of 1.30 for *Sense and Sensibility*.  $\sigma^*$  is greater than  $\sigma$  in four cases, but this ratio is only substantially less than 1 for *Harvey*. The median of the consistent ratios of the shape factor estimates is 1.16 (IQR: 0.10, 95% CI: 1.15, 1.17). Therefore, the average discrepancy between the location estimates is 8% and the average discrepancy between the shape factor estimates is 16%.

This method of assessing goodness-of-fit depends on the proposition that *if* the shot lengths of a film are lognormally distributed *then*  $\frac{G}{M}$  and  $\frac{\sigma^*}{\sigma}$  are approximately equal to 1. However, there are numerous occasions when the two ratios are both approximately equal to one but the normal probability plots and/or the hypothesis tests lead us to reject the lognormal distribution as a model for the data. This may occur if the distribution exhibits bimodality after the transformation has been applied, if the distribution is symmetrical and leptokurtic, or if one of the tails of the distribution deviates markedly from the theoretical distribution while the remainder of the data points are reasonably well fitted. Therefore, the ratios of the geometric mean to the median and of  $\sigma^*$  to  $\sigma$  cannot be considered reliable evidence a shot length distribution *is* lognormally distributed. Relying on these ratios to conclude shots lengths are lognormal leads to logically flawed reasoning by affirming the consequent in the above proposition. These ratios are reliable evidence (by *modus tollens*) the data is not lognormally distributed and where we observe large discrepancies we can be confident the assumed model is not appropriate.

However, we note the case of *Brief Encounter* as a film for which lognormality is a reasonable assumption, as Figure 3a and the results in Table 1 demonstrate, but where the ratios of the location (1.07) and shape (1.08) estimates are greater than many films where lognormality is not a reasonable assumption. The 7% discrepancy between  $G$  and  $M$  for this film is just below the median discrepancy for the whole sample, and relying on this statistic may lead to rejection of lognormality when it is appropriate in this instance. The problem of interpreting these results therefore becomes one of deciding what constitutes a 'large discrepancy' between estimates.

We therefore conclude that it is not possible to accurately determine between lognormal shots length distributions and those that are not by this method alone. Relying on this method may lead researchers to reject the null hypothesis when it should have been accepted (Type I error) or to accept the null hypothesis when it should have been rejected (Type II error). The results of the hypothesis tests in Table 1 indicate Type II errors will be more common and that researchers will fail to reject lognormality as a model for their data by relying on these ratios.

#### 4.2.2 Normal probability plots with log-transformed data

Figure 3 presents the normal probability plots for the log-transformed shot length data of eight films. If the log-transformed data is well fitted by a normal distribution with parameters  $\mu$  and  $\sigma$  then the data points will lie along a straight line. *Brief Encounter* (Figure

3a) and *Barry Lyndon* (Figure 3b) both demonstrate this pattern, and we conclude that, in light of the other results, the lognormal distribution is an appropriate model for shot lengths in these films. The remaining six plots show that applying a logarithmic transformation does not necessarily produce a symmetric distribution and that there are a variety of reasons for this.

A logarithmic transformation is applied to data in the expectation it will normalise skewed data and remove the influence of outliers. However, it is clear that in some cases such a transformation does not make the data symmetrical or completely solve the problem of outliers. De Long, Brunick, and Cutting provide the example of *A Night at the Opera* to illustrate the lognormality of shot lengths, but it is clear from the probability plot for this film (Figure 3c) that a lognormal distribution is not a good fit. Specifically, the distribution of shot lengths is positively skewed even after the data is log-transformed and the fitted distribution  $\mathcal{N} \sim (1.5607, 1.2716^2)$  underestimates the frequency of shorter shots while overestimating the number of longer takes. There are also outliers evident in both tails after a transformation has been applied. For this film there are discrepancies of 19% between the geometric mean and the median and of 26% for the two shape factor estimates – both substantially above the average values for these ratios. The ‘evidence’ presented by De Long, Brunick, and Cutting in support of their argument is both *categorically wrong* (it is wrong without qualification) and a *category error* (it is an error of classification). A similar pattern can be seen in the plot of *Pretty Woman* (Figure 3g), and the failure to eliminate the positive skew and/or outliers from the data is common right across the time period covered.

*Harvey* (Figure 3d) evidently has a leptokurtic distribution with heavier tails than expected. This means that although the distribution of  $\log(X)$  for this film is roughly symmetrical it has a higher peak than a normal distribution, indicating that the log-transformed shot lengths for this film exhibit less variation than expected (and thereby explaining why the ratio of shape factors was less than 1 for this film) with a increased density of shot lengths in both tails. A lognormal distribution clearly cannot accurately model this distribution and it is important to consider not only the skew but also the kurtosis in interpreting shot lengths distributions.

The normal probability plot for *The Apartment* (Figure 3e) appears to be similar to *Brief Encounter* and *Barry Lyndon*, but we conclude the lognormal distribution is not appropriate based on its bimodality after transformation (see below). Bimodality is evident in Figure 3e as the distribution jumps from below the fitted line to the predicted values, but it is far easier to this pattern in the histogram of the transformed data in Figure 4. We can be confident that when the data points deviate substantially from a straight line there is strong evidence against lognormality, but in this instance the apparent linearity of the plot may be easily misinterpreted. Again, this indicates that reliance on a single method for assessing goodness-of-fit is dangerous and may be avoided by using normal probability plots with the Shapiro-Francia statistic and histograms of the log-transformed data. This plot also demonstrates a problem with applying a logarithmic transformation: because the logarithmic transformation stretches the interval  $[0, 1]$  and compresses the interval  $[1, \infty]$ , we may find that some data points appear as outliers in the lower tail of the distribution. This can clearly be seen in Figure 3e and is also apparent in Figure 4. The ‘creation’ of outliers in the lower tail of a distribution means that in a handful of case the log-transformed data may exhibit negative skew, but such instances are relatively rare.

*Shampoo* (Figure 3f) exhibits a common pattern in probability plots for films from the 1960s onwards, deviating markedly from the theoretical distribution in its lower tail. This plot indicates that lower tail of the distribution is heavier than would be expected if  $\log(X)$  were

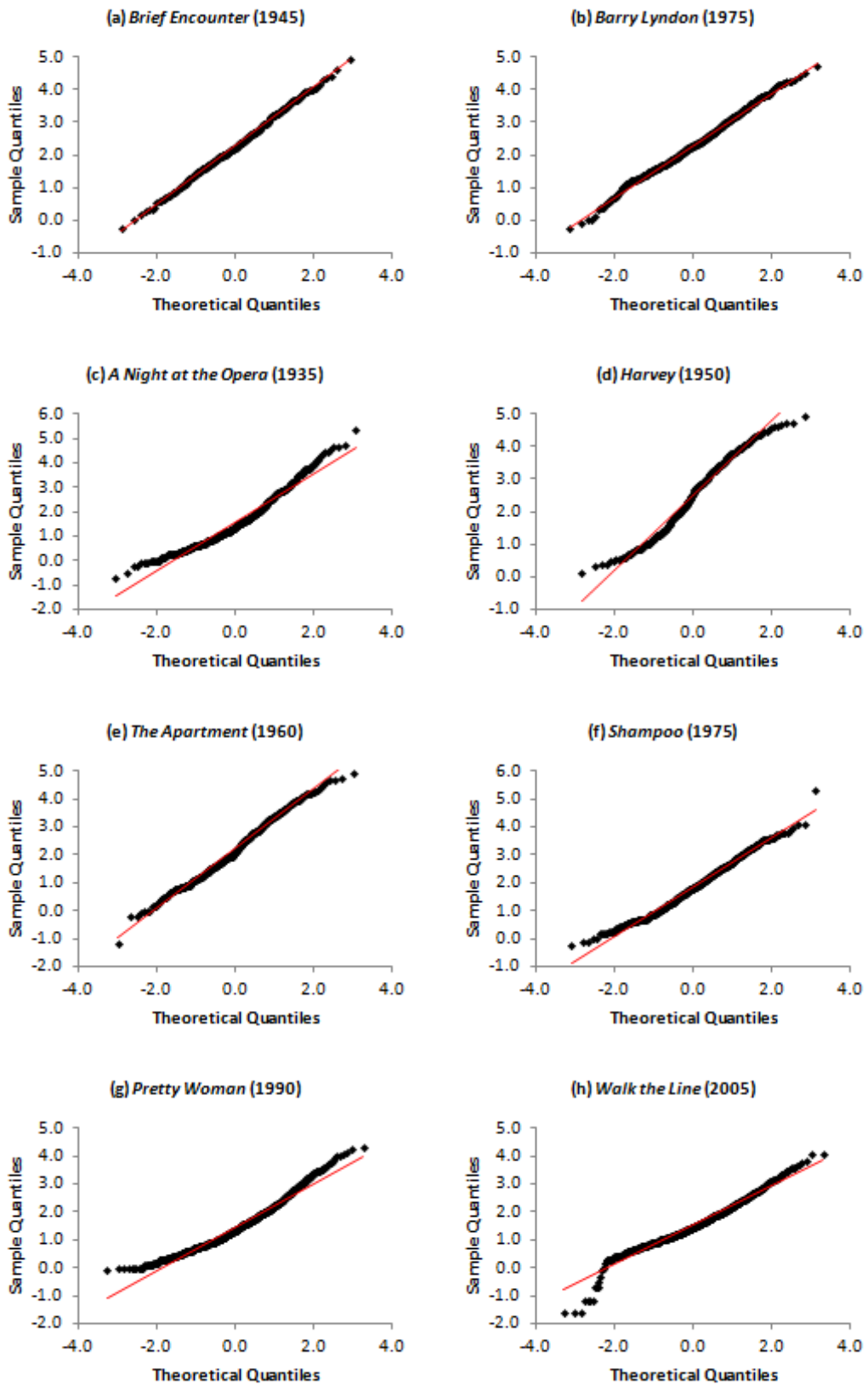


Figure 3 Normal probability plots of log-transformed shot lengths for eight films

normally distributed, and, therefore, that a lognormal distribution underestimates the density of shorter shots in this film. Since this feature is evident for many films in the sample this implies that assuming the lognormal distribution as general distribution of film style will systematically give misleading descriptions of film style. *Shampoo* also includes an outlier in the upper tail: the longest shot for this film is given as 197.9 seconds and is three times greater than the second longest shot (60.3s), and a lognormal distribution simply will not adequately model a data set containing such an extreme outlier. This outlier can be clearly seen to be substantially different after a logarithmic transformation has been applied in the upper right corner of Figure 3f. *Walk the Line* (Figure 3h) shows the opposite pattern to *Shampoo*, with a lighter lower tail than expected indicating that the fitted normal distribution overestimates the density of shorter shots. This pattern is much less common in this sample.

### 4.2.3 The Shapiro-Francia test

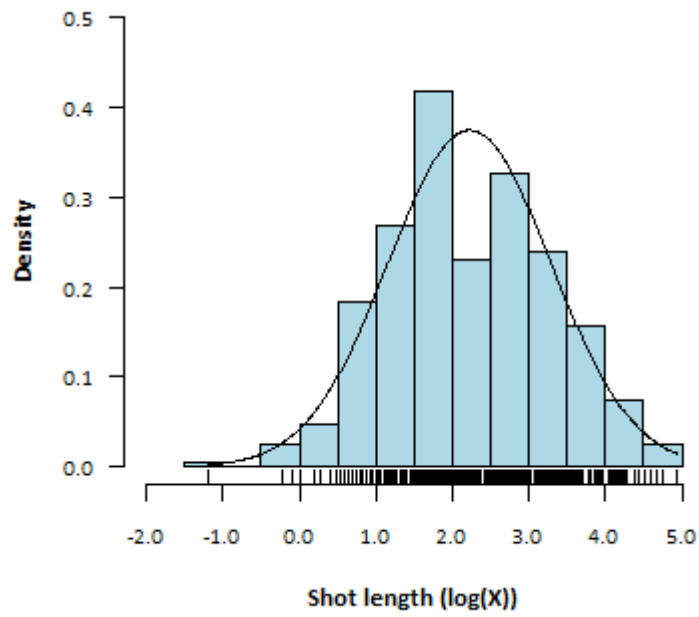
The null hypothesis of lognormality was rejected for 127 of the 134 films tested with the Shapiro-Francia test. This clearly indicates that the assumption of lognormality is not justified for the vast majority of films.

### 4.2.4 The Jarque-Bera test

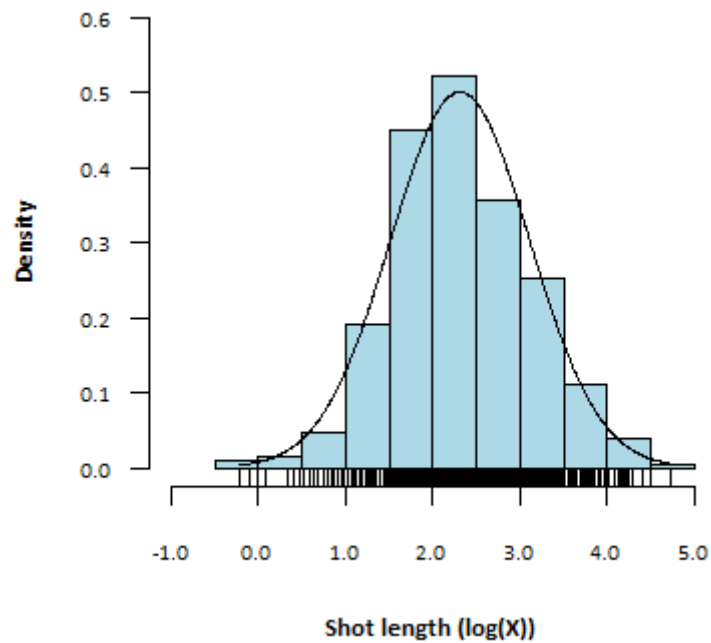
The null hypothesis of lognormality was rejected for 124 of the 134 films tested with Jarque-Bera test, and again we conclude that lognormality is not appropriate in the vast majority of cases.

There is a discrepancy between the results of the SF test and the JB test for *The Apartment* and *Barry Lyndon*. For *The Apartment*, the SF test rejected the null hypothesis while the JB statistic was not statistically significant in this instance. Nonetheless, we consider the lognormal distribution to be inappropriate based on the above average difference between the two location estimates (1.11) and on the histogram of the log-transformed shot lengths (see Figure 4). Although the resulting distribution is symmetrical, the shot lengths in this film are bimodal under a logarithmic transformation and the peak of the fitted distribution lies directly over the trough between the modes.

In the case of *Barry Lyndon*, the SF test again rejected lognormality while the JB test indicates the null hypothesis is a plausible model for this data. There is no large discrepancy between the location estimates (1.03) and the ratio of the shape factor estimates (1.06) is well below the sample median. As noted above, the probability plot for this film indicates a strong linear relationship between the theoretical and observed values, and the histogram of  $\log(X)$  and the fitted normal distribution (Figure 5) also support the interpretation that a lognormal distribution is a reasonable assumption for this data.



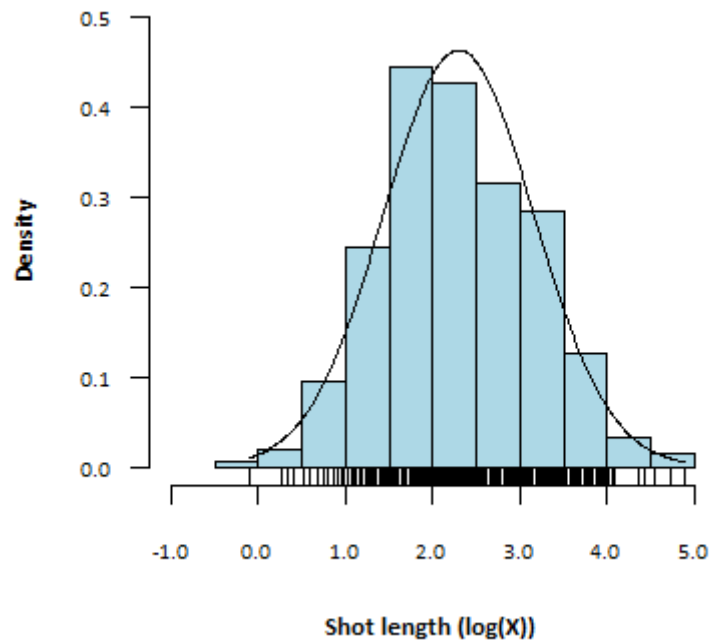
**Figure 4** Histogram of log-transformed shot lengths in *The Apartment* with fitted normal distribution  $\mathcal{N} \sim (2.2241, 1.0642^2)$



**Figure 5** Histogram of log-transformed shot lengths in *Barry Lyndon* with fitted normal distribution  $\mathcal{N} \sim (2.3138, 0.7965^2)$

4.2.5 *Blood on the Sun*

There is one case where the results are inconclusive. In the case of *Blood on the Sun*, the null hypothesis was rejected for the SF test but not for the JB test. The histogram of  $\log(X)$  and the fitted normal distribution for this film are similarly inconclusive: although the two density estimates in Figure 6 appear to be similar there is a greater than average discrepancy in its location estimates (1.10) and the ratio between the shape factor estimates (1.13) is below average but still much greater than 1. Thus we find some evidence against lognormality for this film and some for, and although we cannot quite reject the null hypothesis of lognormality this does not justify any such assumption for shot lengths in this film. In fact, given these results are inconclusive we caution against making any such assumption and recommend a conservative approach and the use of robust methods, particularly in light of the size of the ratios of the descriptive statistics.



**Figure 6** Histogram of log-transformed shot lengths in *Blood on the Sun* with fitted normal distribution  $\mathcal{N} \sim (2.2992, 0.8624^2)$

4.3 Discussion

The first stage in the statistical analysis of film style should always be a detailed examination of the data. *Exploratory data analysis* (EDA) is an approach to data analysis characterized by scepticism of methods that may obscure the structure of the data and openness to unanticipated patterns (Tukey 1977, Hartwig & Dearing 1979, Behrens 1997). EDA employs a range of methods to maximise insight into a data set by revealing the underlying structure of the data and extracting the relevant features, identifying outliers, generating hypotheses, and testing underlying statistical assumptions. EDA places substantial emphasis on resistant and robust methods requiring few assumptions about the data and which are applicable in a wide range of circumstances: ‘Robustness is important in EDA because the underlying form of the data cannot always be presumed, and statistics that can be easily fooled (like the

mean) may mislead' (Behrens 1997: 143). The EDA approach is clearly different to *classical* statistics in which models and hypotheses are determined before seeing the data, and in which the objectives of statistical analysis are the estimation of model parameters and to confirm the presence or absence of specific features of the data. EDA does not replace confirmatory hypothesis tests but assumes that the more we know about our data the more effective our subsequent analyses will be. Exploratory data analysis is then a fundamental part of statistical research, and any study 'that does not include a thorough exploratory data analysis is not complete' (Kundzewicz & Robson 2004: 9).

A fundamental principle of EDA is that obtaining a well-rounded view of a data set requires the use of a range of exploratory methods in conjunction, combining tabular, numerical, and graphical representations in a flexible and intuitive manner. This paper has demonstrated the importance of using a range of different numerical and graphical exploratory techniques alongside normality tests. It is clear that there is no single method for judging the fit of a lognormal distribution to shot length data, and it is necessary to use multiple methods to properly evaluate the assumptions that underpin the statistical analysis of film style. It is also clear that comparing the density trace of a fitted lognormal distribution to histograms of the untransformed data is misleading and should not be used. The histogram of the log-transformed data with a fitted normal distribution is more informative, and, when used alongside several descriptive statistics and normal probability plots, allows us to gain some insight into the results of the formal hypothesis tests of normality.

We tested the assumption the lognormal distribution for 134 Hollywood films released from 1935 to 2005 using a range of exploratory and confirmatory statistical methods, and we found this assumption was not justified in 125 cases. Designing a study of film style based on this assumption will lead to flawed inferences due to incorrect descriptions of film style. Furthermore, reliance of parametric statistical tests that assume data is normally distributed (after an appropriate transformation is applied) may result in a loss of statistical power. Departures from the underlying normal distribution may be arbitrarily small but still lead to fundamentally flawed conclusions (Wilcox 1998, Huber & Ronchetti 2009: 1-2). From the above results we see those departures take such a variety of forms – bimodality, positive skew, negative skew, deviations in the lower or upper tails, outliers in the lower and/or upper tails, leptokurtosis – that it is unlikely there is a simple and general method for dealing with these deviations.

A far simpler approach that avoids these problems is to make no such assumption, and to use statistical methods that do not depend on assumptions of lognormality. This can be illustrated in the choice of measures of central tendency and dispersion for describing shot length distributions.

Salt has previously claimed the ASL is a useful statistic of film style, but this appears inconsistent with claims the data is lognormally distributed. It would seem obvious that we should give up the mean as a statistic of film style since it does not locate the centre of the data for some other measure of central tendency such as the geometric mean or the median. Nonetheless, he advocates retaining the mean shot length even though the distribution of shot lengths is highly skewed though it is not clear what the mean shot length means in this context since it no longer functions as a measure of central tendency. Statements such as 'the mean shot length of *Little Annie Rooney* (minus titles) is 4.6 seconds' no longer mean what the majority of film scholars think they mean as this statistic no longer intended to be used as a description of the cutting rate of the film. However, this argument is invalid since it is



based the relationship between the arithmetic mean, the median, and  $\sigma$  under lognormality and this fundamental assumption is not justified in general.

The justification given by De Long, Brunick, and Cutting for preferring the median shot length to the mean because shot lengths are lognormal is clearly flawed as the premise of this argument is invalid for the vast majority of cases. The median is a superior measure of central tendency not because we are justified in assuming lognormality but because it locates the centre of a distribution *irrespective* of its shape. The proper justification for using the median shot length to describe the style of a film is that it is *resistant* to the effects of outliers and *robust* to deviations from the assumed model due to its high breakdown point and bounded influence function (Wilcox 1998, 2005). Neither the arithmetic nor geometric means are resistant or robust and use of these statistics lead to flawed interpretations of the data. Additionally, the shape factor of the log-transformed shot length data is not a resistant or robust measure of dispersion and similarly leads to incorrect conclusions. Robust measures of dispersion such as the interquartile range,  $S_n$ , or  $Q_n$  (Rousseeuw & Croux 1993) are far superior to  $\sigma$  or  $\sigma^*$ , and this is again due to their high breakdown points and bounded influence functions. Calculating these statistics from the untransformed data expresses the dispersion of shot lengths in seconds making the interpretation of film style easier than using  $\sigma$  with the ASL, and they have clearly defined meanings.

### 5. Conclusion

This paper examined the claim that lognormality is a reasonable assumption for shot length distributions of Hollywood films. In reviewing the claims put forward to support this assumption we find them lacking in methodological detail and statistical rigour, and certainly unable to justify the conclusions presented. In testing the shot length data of Hollywood films we find that, while the lognormal distribution is an adequate model for a handful of films, this is not the case for the vast majority of films in the sample. Consequently, we conclude there is no justification for the claim that the lognormal distribution is an appropriate parametric model for the shot length distributions of Hollywood films.

### References

- Aitchison J and Brown JAC** 1957 *The Lognormal Distribution, with Special Reference to its Use in Economics*. Cambridge: Cambridge University Press.
- Behrens JT** 1997 Principles and practices of exploratory data analysis, *Psychological Methods* 2 (2): 131-160.
- Burmester DE and Hull DA** 1997 Using lognormal distributions and lognormal probability plots in probabilistic risk assessments, *Human and Ecological Risk Assessment: An International Journal* 3 (2): 235-255.
- Cullen AC and Frey HC** 1999 *Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. New York: Plenum.
- D'Agostino RB** 1986 Tests for the normal distribution, in RB D'Agostino and MA Stephens (eds.) *Goodness of Fit Techniques*. New York: Marcel Dekker: 367-419.
- De Long J, Brunick KL, and Cutting JE** 2012 Film through the human visual system: finding patterns and limits, in JC Kaufman and DK Simonton (eds.) *The Social Science of Cinema*. New York: Oxford University Press: in press. An online version of this paper is available at

<http://people.psych.cornell.edu/~jec7/pubs/socialsciencecinema.pdf>, accessed 18 January 2012.

**Filliben JJ** 1975 The probability plot correlation coefficient test for normality, *Technometrics* 17 (1): 111-117.

**Hartwig F and Dearing BE** 1979 *Exploratory Data Analysis*. Newbury Park, CA: Sage.

**Huber PJ and Ronchetti EM** 2009 *Robust Statistics*, second edition. New York: John Wiley & Sons.

**Jarque CM and Bera AK** 1987 A test for normality of observations and regression residuals, *International Statistical Review* 55 (2): 163-172.

**Kleiber C and Kotz S** 2003 *Statistical Size Distributions in Economics and Actuarial Sciences*. Hoboken, NJ: John Wiley & Sons.

**Kundzewicz ZB and Robson AJ** 2004 Change detection in hydrological records: a review of the methodology, *Hydrological Sciences* 49 (1): 7-19.

**Limpert E, Stahel WA, and Abbt M** 2001 Log-normal distributions across the sciences: keys and clues, *Bioscience* 51 (5): 341-352.

**Looney SW and Gulledge TR** 1985 Use of the correlation coefficient with normal probability plots, *The American Statistician* 39 (1): 75-79.

**Quinn GP and Keogh MJ** 2002 *Experimental Design and Data Analysis for Biologists*. Cambridge: Cambridge University Press.

**Rousseeuw PJ and Croux C** 1993 Alternatives to the median absolute deviation, *Journal of the American Statistical Association* 88: 1273-1283.

**Shapiro SS and Francia RS** 1972 An approximate analysis-of-variance test for normality, *Journal of the American Statistical Association* 67: 215-216.

**Salt B** 2006 *Moving into Pictures: More on Film History, Style, and Analysis*. London: Starword.

**Salt B** 2011 The metrics in Cinematics, [http://www.cinematics.lv/metrics\\_in\\_cinematics.php](http://www.cinematics.lv/metrics_in_cinematics.php), accessed 27 January 2011.

**Thode HC** 2002 *Testing for Normality*. New York: Marcel Dekker, Inc.

**Tukey JW** 1977 *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

**Wilcox RR** 1998 How many discoveries have been lost by ignoring modern statistical methods?, *American Psychologist* 53 (3): 300-314.

**Wilcox RR** 2005 *Introduction to Robust Estimation and Hypothesis Testing*, second edition. Burlington, MA: Elsevier Academic Press.

## The lognormal distribution and Hollywood cinema

**Table 1** Results of statistical tests of the null hypothesis shot lengths in Hollywood films are lognormally distributed (see text for discussion of films marked \*)

Title	Year	Median (M)	Geometric mean (G)	$\frac{G}{M}$	$\sigma$	$\sigma^*$	$\frac{\sigma^*}{\sigma}$	Shapiro-Francia	p	Jarque-Bera	p	Lognormal?
A Tale of Two Cities	1935	5.2	5.7	1.09	0.9195	1.0531	1.15	0.9852	<0.01	22.88	<0.01	NO
Top Hat	1935	5.4	6.1	1.13	0.9428	1.1507	1.22	0.9677	<0.01	48.67	<0.01	NO
Les Miserables	1935	5.0	5.4	1.07	0.8795	0.9722	1.11	0.9924	<0.01	11.04	<0.01	NO
The Informer	1935	7.4	7.4	1.00	0.8450	0.8614	1.02	0.9928	0.02	6.48	0.04	NO
Westward Ho	1935	4.0	4.4	1.10	0.8674	1.0011	1.15	0.9866	<0.01	12.75	<0.01	NO
A Night at the Opera	1935	4.0	4.8	1.19	1.0081	1.2716	1.26	0.9585	<0.01	62.48	<0.01	NO
Anna Karenina	1935	6.5	6.4	0.99	0.8651	0.8712	1.01	0.9928	0.01	7.02	0.03	NO
The 39 Steps	1935	4.1	4.9	1.20	0.9864	1.2730	1.29	0.9674	<0.01	44.95	<0.01	NO
Captain Blood	1935	4.4	4.4	0.99	0.7959	0.7876	0.99	0.9977	0.10	0.27	0.88	YES
Mutiny on the Bounty	1935	4.5	4.9	1.10	0.9677	1.0647	1.10	0.9934	<0.01	13.06	<0.01	NO
Fantasia	1940	6.4	6.7	1.04	0.9570	1.0201	1.07	0.9963	0.10	4.04	0.13	YES
Foreign Correspondent	1940	4.3	4.6	1.07	0.8936	1.0097	1.13	0.9852	<0.01	31.15	<0.01	NO
Grapes Of Wrath	1940	6.8	6.9	1.02	0.8554	0.9073	1.06	0.9891	<0.01	11.88	<0.01	NO
Pinocchio	1940	4.0	4.2	1.05	0.7315	0.7996	1.09	0.9939	<0.01	10.19	<0.01	NO
Rebecca	1940	4.8	5.7	1.18	0.9221	1.1599	1.26	0.9651	<0.01	64.00	<0.01	NO
Santa Fe Trail	1940	4.3	4.5	1.05	0.8847	0.9367	1.06	0.9932	<0.01	12.85	<0.01	NO
The Great Dictator	1940	9.2	9.2	1.00	0.9992	0.9878	0.99	0.9976	0.50	2.45	0.29	YES
The Letter	1940	7.1	8.1	1.14	1.0146	1.1677	1.15	0.9888	<0.01	8.76	0.01	NO
Thief Of Bagdad	1940	3.6	3.9	1.08	0.7471	0.8750	1.17	0.9883	<0.01	30.03	<0.01	NO
Bells Of St Marys	1945	6.0	6.9	1.15	0.9059	1.0972	1.21	0.9639	<0.01	40.62	<0.01	NO
Blood On The Sun	1945	9.1	10.0	1.10	0.8624	0.9745	1.13	0.9936	0.05	5.11	0.08	N/A*
Brief Encounter	1945	9.2	9.8	1.07	0.8927	0.9632	1.08	0.9982	0.95	0.42	0.81	YES
Detour	1945	12.3	12.6	1.02	0.8968	0.8992	1.00	0.9893	0.04	6.77	0.03	NO
In Pursuit To Algiers	1945	5.8	6.9	1.19	1.0229	1.1888	1.16	0.9754	<0.01	13.7	<0.01	NO
Leave Her to Heaven	1945	5.9	7.0	1.18	0.8700	1.0983	1.26	0.9587	<0.01	45.02	<0.01	NO
Lost Weekend	1945	8.2	8.3	1.01	0.8893	0.8870	1.00	0.9894	<0.01	10.88	<0.01	NO
Spellbound	1945	5.1	5.8	1.14	1.0242	1.1970	1.17	0.9763	<0.01	27.87	<0.01	NO

## The lognormal distribution and Hollywood cinema

Title	Year	Median (M)	Geometric mean (G)	$\frac{G}{M}$	$\sigma$	$\sigma^*$	$\frac{\sigma^*}{\sigma}$	Shapiro-Francia	p	Jarque-Bera	p	Lognormal?
All About Eve	1950	4.9	5.8	1.17	1.0199	1.2441	1.22	0.9581	<0.01	61.82	<0.01	NO
Annie Get Your Gun	1950	5.0	6.1	1.22	1.2698	1.4607	1.15	0.9702	<0.01	20.65	<0.01	NO
Born Yesterday	1950	6.9	7.9	1.15	1.2012	1.3356	1.11	0.9796	<0.01	12.39	<0.01	NO
Cheaper by the Dozen	1950	7.3	8.0	1.10	1.0143	1.0923	1.08	0.9868	<0.01	9.00	<0.01	NO
Cinderella	1950	3.1	3.2	1.02	0.6724	0.7255	1.08	0.9884	<0.01	45.19	<0.01	NO
Harvey	1950	13.2	12.3	0.93	1.1532	1.0242	0.89	0.9763	<0.01	12.07	<0.01	NO
The Asphalt Jungle	1950	6.6	7.2	1.09	0.8304	0.9799	1.18	0.9771	<0.01	30.80	<0.01	NO
The Flame And The Arrow	1950	4.0	4.3	1.08	0.8824	1.0040	1.14	0.9887	<0.01	21.59	<0.01	NO
Battle Cry	1955	6.0	6.5	1.08	0.8610	0.9798	1.14	0.9897	<0.01	20.96	<0.01	NO
East Of Eden	1955	5.1	5.9	1.16	0.9547	1.1556	1.21	0.9682	<0.01	41.51	<0.01	NO
Lady And The Tramp	1955	3.8	3.9	1.02	0.6581	0.6945	1.06	0.9943	<0.01	8.73	0.01	NO
Mr Roberts	1955	6.5	7.5	1.16	0.8856	1.1220	1.27	0.9688	<0.01	57.6	<0.01	NO
Night of the Hunter	1955	4.9	5.0	1.03	0.8837	0.9165	1.04	0.9965	0.10	4.51	0.11	YES
Rebel Without A Cause	1955	4.8	5.2	1.09	0.8385	0.9821	1.17	0.9825	<0.01	34.75	<0.01	NO
Seven Year Itch	1955	11.3	13.3	1.18	1.1617	1.2973	1.12	0.9755	<0.01	9.10	0.01	NO
The Ladykillers	1955	5.1	5.2	1.03	0.9100	0.9317	1.02	0.9946	0.02	7.72	0.02	NO
The Trouble With Harry	1955	5.1	5.5	1.08	0.8806	1.0121	1.15	0.9809	<0.01	27.19	<0.01	NO
Butterfield 8	1960	6.1	7.1	1.16	0.9697	1.1330	1.17	0.9745	<0.01	23.12	<0.01	NO
Exodus	1960	13.8	13.6	0.99	0.9903	0.9828	0.99	0.9957	0.12	4.67	0.10	YES
Inherit The Wind	1960	7.5	8.4	1.12	1.0069	1.1875	1.18	0.9784	<0.01	27.08	<0.01	NO
Magnificent Seven	1960	4.5	4.8	1.07	0.8179	0.9345	1.14	0.9907	<0.01	24.13	<0.01	NO
Ocean's 11	1960	7.4	8.0	1.07	0.8801	1.0061	1.14	0.9790	<0.01	25.97	<0.01	NO
Peeping Tom	1960	5.2	5.6	1.08	0.9499	1.0695	1.13	0.9877	<0.01	16.23	<0.01	NO
Spartacus	1960	4.9	5.4	1.11	0.8587	1.0267	1.20	0.9823	<0.01	57.13	<0.01	NO
Swiss Family Robinson	1960	3.1	3.3	1.07	0.7781	0.8910	1.15	0.9890	<0.01	38.31	<0.01	NO
The Apartment	1960	8.3	9.2	1.11	1.0642	1.1496	1.08	0.9915	0.01	5.51	0.06	NO*
Time Machine	1960	3.4	3.9	1.15	0.8773	1.0813	1.23	0.9713	<0.01	61.20	<0.01	NO

## The lognormal distribution and Hollywood cinema

Title	Year	Median (M)	Geometric mean (G)	$\frac{G}{M}$	$\sigma$	$\sigma^*$	$\frac{\sigma^*}{\sigma}$	Shapiro-Francia	p	Jarque-Bera	p	Lognormal?
Dr Zhivago	1965	5.9	6.4	1.08	0.7796	0.9035	1.16	0.9851	<0.01	38.26	<0.01	NO
Flight Phoenix	1965	3.3	3.6	1.09	0.7312	0.8863	1.21	0.9809	<0.01	84.52	<0.01	NO
Flying Machines	1965	3.9	4.4	1.12	0.8574	1.0759	1.25	0.9613	<0.01	139.55	<0.01	NO
Help	1965	2.6	2.8	1.07	0.7054	0.8142	1.15	0.9913	<0.01	26.22	<0.01	NO
Shenandoah	1965	4.0	4.3	1.08	0.7749	0.9068	1.17	0.9811	<0.01	38.24	<0.01	NO
Sound Of Music	1965	4.2	4.5	1.08	0.8606	0.9981	1.16	0.9796	<0.01	69.15	<0.01	NO
That Darn Cat	1965	3.4	3.8	1.10	0.7419	0.8940	1.21	0.9852	<0.01	41.27	<0.01	NO
The Great Race	1965	4.0	4.4	1.11	0.9300	1.0809	1.16	0.9841	<0.01	42.96	<0.01	NO
Thunderball	1965	2.5	2.8	1.11	0.7821	0.9517	1.22	0.9759	<0.01	114.85	<0.01	NO
What's New Pussycat	1965	4.2	4.9	1.17	0.9720	1.1943	1.23	0.9645	<0.01	69.98	<0.01	NO
Airport	1970	4.7	5.5	1.17	0.7921	1.0037	1.27	0.9787	<0.01	44.48	<0.01	NO
Aristocats	1970	3.3	3.4	1.02	0.5860	0.6203	1.06	0.9976	0.08	1.19	0.55	YES
Beneath The Planet Of The Apes	1970	2.8	2.9	1.03	0.9785	0.9944	1.02	0.9716	<0.01	136.79	<0.01	NO
Catch 22	1970	5.6	6.2	1.10	1.1576	1.3137	1.13	0.9804	<0.01	22.68	<0.01	NO
Five Easy Pieces	1970	3.6	4.2	1.17	0.9476	1.2092	1.28	0.9568	<0.01	103.62	<0.01	NO
Kelly's Heroes	1970	4.1	4.3	1.05	0.8422	0.9094	1.08	0.9951	<0.01	8.78	0.01	NO
Patton	1970	5.0	5.6	1.12	0.9312	1.0964	1.18	0.9865	<0.01	32.34	<0.01	NO
Tora! Tora! Tora!	1970	4.5	4.8	1.07	0.9010	0.9937	1.10	0.9948	<0.01	7.72	0.02	NO
Barry Lyndon	1975	9.8	10.1	1.03	0.7965	0.8415	1.06	0.9960	0.05	1.13	0.57	YES*
Three Days of the Condor	1975	3.4	3.7	1.08	0.9017	1.0258	1.14	0.9875	<0.01	33.54	<0.01	NO
One Flew Over the Cuckoo's Nest	1975	3.6	3.8	1.06	0.7439	0.8372	1.13	0.9911	<0.01	28.98	<0.01	NO
Dog Day Afternoon	1975	3.1	3.4	1.10	0.8066	0.9926	1.23	0.9633	<0.01	165.00	<0.01	NO
Jaws	1975	3.6	4.0	1.12	0.9254	1.1230	1.21	0.9788	<0.01	74.3	<0.01	NO
The Man Who Would Be King	1975	4.9	5.4	1.09	0.8352	0.9755	1.17	0.9799	<0.01	38.06	<0.01	NO
Monty Python The Holy Grail	1975	2.6	3.0	1.14	0.9279	1.1537	1.24	0.9665	<0.01	101.45	<0.01	NO
Return Of The Pink Panther	1975	3.6	4.2	1.14	1.0955	1.2814	1.17	0.9726	<0.01	42.18	<0.01	NO
The Rocky Horror Picture Show	1975	3.3	3.6	1.08	0.9359	1.0941	1.17	0.9838	<0.01	53.68	<0.01	NO
Shampoo	1975	6.1	6.3	1.04	0.8805	0.9413	1.07	0.9896	<0.01	10.43	<0.01	NO

## The lognormal distribution and Hollywood cinema

Title	Year	Median (M)	Geometric mean (G)	$\frac{G}{M}$	$\sigma$	$\sigma^*$	$\frac{\sigma^*}{\sigma}$	Shapiro-Francia	p	Jarque-Bera	p	Lognormal?
Airplane	1980	4.3	4.8	1.11	0.8588	1.0033	1.17	0.9819	<0.01	19.99	<0.01	NO
Coal Miner's Daughter	1980	5.3	6.2	1.18	0.9343	1.1368	1.22	0.9716	<0.01	35.94	<0.01	NO
The Empire Strikes Back	1980	2.9	3.0	1.04	0.7588	0.8436	1.11	0.9918	<0.01	35.53	<0.01	NO
Nine To Five	1980	3.9	4.1	1.05	0.8418	0.9338	1.11	0.9899	<0.01	27.47	<0.01	NO
Ordinary People	1980	3.5	3.8	1.09	0.9074	1.0748	1.18	0.9753	<0.01	79.57	<0.01	NO
Popeye	1980	3.3	3.5	1.07	0.7808	0.9020	1.16	0.9860	<0.01	51.95	<0.01	NO
Stir Crazy	1980	4.4	4.8	1.09	0.7976	0.9423	1.18	0.9814	<0.01	41.04	<0.01	NO
Superman 2	1980	2.5	2.7	1.09	0.7920	0.9473	1.20	0.9818	<0.01	96.57	<0.01	NO
The Blue Lagoon	1980	3.0	3.4	1.12	0.7910	0.9754	1.23	0.9733	<0.01	91.34	<0.01	NO
Urban Cowboy	1980	3.5	4.0	1.13	0.8212	1.0140	1.23	0.9772	<0.01	79.56	<0.01	NO
Back To The Future	1985	2.7	3.1	1.13	0.9090	1.1025	1.21	0.9720	<0.01	86.73	<0.01	NO
Cocoon	1985	3.9	4.2	1.08	0.7757	0.9108	1.17	0.9836	<0.01	44.10	<0.01	NO
Out Of Africa	1985	3.5	3.7	1.05	0.7962	0.8786	1.10	0.9932	<0.01	16.93	<0.01	NO
Police Academy 2	1985	3.0	3.3	1.12	0.8654	1.0370	1.20	0.9753	<0.01	52.22	<0.01	NO
Rambo II	1985	2.0	2.2	1.08	0.7457	0.8984	1.20	0.9757	<0.01	145.18	<0.01	NO
Spies Like Us	1985	2.5	2.7	1.09	0.7703	0.9259	1.20	0.9851	<0.01	66.54	<0.01	NO
The Color Purple	1985	4.7	5.1	1.08	0.8471	0.9672	1.14	0.9930	<0.01	20.44	<0.01	NO
Witness	1985	4.2	4.4	1.04	0.8322	0.9059	1.09	0.9945	<0.01	12.12	<0.01	NO
Dick Tracy	1990	2.8	2.9	1.03	0.7803	0.8474	1.09	0.9928	<0.01	21.91	<0.01	NO
Die Hard 2	1990	2.1	2.2	1.07	0.7519	0.8709	1.16	0.9837	<0.01	81.04	<0.01	NO
Ghost	1990	3.4	3.6	1.05	0.8156	0.9452	1.16	0.9862	<0.01	87.77	<0.01	NO
Goodfellas	1990	4.2	4.4	1.06	0.9026	1.0037	1.11	0.9920	<0.01	21.94	<0.01	NO
Home Alone	1990	3.1	3.2	1.03	0.7726	0.8306	1.08	0.9944	<0.01	15.30	<0.01	NO
Hunt For Red October	1990	4.7	5.0	1.06	0.8007	0.9009	1.13	0.9915	<0.01	17.14	<0.01	NO
Pretty Woman	1990	3.8	4.3	1.13	0.7886	0.9896	1.25	0.9675	<0.01	103.64	<0.01	NO
Teenage Mutant Ninja Turtles	1990	2.8	3.1	1.12	0.8028	0.9906	1.23	0.9800	<0.01	72.27	<0.01	NO
Total Recall	1990	2.4	2.6	1.07	0.7956	0.9190	1.16	0.9871	<0.01	48.64	<0.01	NO

## The lognormal distribution and Hollywood cinema

Title	Year	Median (M)	Geometric mean (G)	$\frac{G}{M}$	$\sigma$	$\sigma^*$	$\frac{\sigma^*}{\sigma}$	Shapiro-Francia	p	Jarque-Bera	p	Lognormal?
Ace Ventura 2	1995	2.7	3.0	1.11	0.8634	1.0480	1.21	0.9782	<0.01	78.94	<0.01	NO
Apollo 13	1995	3.5	3.7	1.05	0.7240	0.8167	1.13	0.9933	<0.01	27.13	<0.01	NO
Batman Forever	1995	2.4	2.6	1.07	0.8067	0.9410	1.17	0.9806	<0.01	97.64	<0.01	NO
Casper	1995	4.1	4.3	1.04	0.9107	0.9712	1.07	0.9941	<0.01	9.45	<0.01	NO
Goldeneye	1995	2.3	2.4	1.06	0.8182	0.9480	1.16	0.9759	<0.01	132.94	<0.01	NO
Jumanji	1995	2.5	2.6	1.04	0.7536	0.8326	1.10	0.9891	<0.01	41.05	<0.01	NO
Pocahontas	1995	2.8	2.8	1.02	0.6922	0.7304	1.06	0.9955	<0.01	10.38	<0.01	NO
Sense and Sensibility	1995	3.8	4.4	1.15	0.8068	1.0484	1.30	0.9611	<0.01	187.72	<0.01	NO
Toy Story	1995	2.1	2.3	1.10	0.6932	0.8412	1.21	0.9854	<0.01	44.00	<0.01	NO
Castaway	2000	4.5	5.2	1.15	1.0086	1.2186	1.21	0.9758	<0.01	51.43	<0.01	NO
Charlie's Angels	2000	2.0	2.1	1.07	0.7854	0.9282	1.18	0.9863	<0.01	103.71	<0.01	NO
Dinosaur	2000	2.8	3.0	1.06	0.5758	0.6882	1.20	0.9922	<0.01	24.72	<0.01	NO
Erin Brockovich	2000	4.2	4.4	1.04	0.6740	0.7588	1.13	0.9917	<0.01	42.30	<0.01	NO
The Grinch Who Stole Christmas	2000	2.6	2.9	1.10	0.7762	0.9297	1.20	0.9823	<0.01	61.51	<0.01	NO
Scary Movie	2000	2.3	2.6	1.13	0.7759	0.9822	1.27	0.9699	<0.01	127.03	<0.01	NO
The Perfect Storm	2000	3.3	3.6	1.09	0.7257	0.8690	1.20	0.9870	<0.01	54.99	<0.01	NO
What Women Want	2000	2.5	2.8	1.11	0.8163	0.9984	1.22	0.9783	<0.01	136.12	<0.01	NO
X-men	2000	2.0	2.1	1.06	0.7284	0.8343	1.15	0.9892	<0.01	56.33	<0.01	NO
Hitch	2005	2.8	3.0	1.06	0.6983	0.8103	1.16	0.9878	<0.01	62.24	<0.01	NO
King Kong	2005	2.6	2.7	1.05	0.6649	0.7519	1.13	0.9951	<0.01	38.53	<0.01	NO
The Longest Yard	2005	2.3	2.5	1.07	0.6462	0.7843	1.21	0.9775	<0.01	175.94	<0.01	NO
Madagascar	2005	3.0	3.0	0.99	0.8357	0.8369	1.00	0.9950	<0.01	6.61	0.04	NO
Mr & Mrs Smith	2005	2.8	3.0	1.06	0.7966	0.8816	1.11	0.9930	<0.01	21.05	<0.01	NO
Walk The Line	2005	4.3	4.6	1.07	0.7048	0.8121	1.15	0.9733	<0.01	140.96	<0.01	NO
Wedding Crashers	2005	2.4	2.5	1.04	0.7311	0.7962	1.09	0.9954	<0.01	20.62	<0.01	NO

Due to data sets containing shot lengths less than or equal to 0.0 seconds a total of 16 films were excluded from the sample: *Philadelphia Story* (1940), *Anchors Aweigh* (1945), *Mildred Pierce* (1945), *King Solomon's Mines* (1950), *Sunset Boulevard* (1950), *To Catch a Thief* (1955), *Little Big Man* (1970), *Mash* (1970), *Jewel of the Nile* (1985), *Rocky IV* (1985), *Dances with Wolves* (1990), *The Usual Suspects* (1995), *Mission Impossible 2* (2000), *Chicken Little* (2005), *Harry Potter and the Goblet of Fire* (2005), and *Star Wars: Episode III – The Revenge of the Sith* (2005).