Horseshoes, handgrenades, and model fitting: the lognormal distribution is a pretty good model for shot-length distribution of Hollywood films

Jordan DeLong
Cornell University, USA

Abstract

In an article published in Literary and Linguistic Computing, Redfern argues against the use of a lognormal distribution and summarizes previous work as 'lacking in methodological detail and statistical rigour'. This response will summarize the article's methodology and conclusion, arguing that while Redfern finds that films are not 'perfectly' lognormal, this is hardly evidence worthy of the ultimate conclusion that a lognormal fit is 'inappropriate'. Perfection is fleeting, and cannot be expected when modeling real data. Reanalysis of Redfern's methodology and findings shows that the lognormal distribution offers a pretty good fit.

Correspondence:
Jordan DeLong, Department of Psychology, Cornell University, 206 Uris Hall, Ithaca, NY 14853-7601
Email:
jed245@cornell.edu

1 Introduction

In an article recently published in Literary and Linguistic Computing, Redfern (2015) argues against the use of a lognormal distribution when modeling the statistics of Hollywood Film. Redfern reviews three publications that have suggested the lognormal distribution for shot-length distribution including two by Salt (2006, 2011) and a chapter I wrote to appear soon in 'The Social Science of Cinema', a book intended to provide an interdisciplinary introduction to some of the important and diverse work being conducted on film (DeLong *et al.*, 2013). Despite being a chapter dedicated to summarizing the first several years of film-based research conducted in our laboratory, Redfern's criticism revolves around a largely peripheral comment that 'the distribution of shot lengths isn't a normal bell curve, but rather a highly skewed, lognormal distribution'. That quote is from a previous draft and not accurate, as the published version of the chapter reads that the distribution of shot lengths is an 'approximately lognormal distribution'.

This single point caused the chapter to be characterized by Redfern as 'lacking in methodological detail and statistical rigour, and certainly unable to justify the conclusions presented'. Redfern's negative portrayal of a general-audience in-press review chapter was surprising. This commentary will be used as an opportunity to speak more in depth about fitting lognormal probability distribution to data in film and perhaps temper Redfern's concerns.

2 Recap and Review of Redfern

In his article, Redfern (2015) uses a relatively simple methodology to assess whether individual films

Digital Scholarship in the Humanities, Vol. 30, No. 1, 2015.

doi:10.1093/llc/fqt030 Advance Access published on 18 June 2013

follow a strict lognormal distribution. To test his hypothesis that Redfern used shot-length data for 134 films from the database collected for a previous article (Cutting *et al.*, 2010) and disturbed on the Cinemetrics website: http://www.cinemetrics.lv. The list of shot lengths for each individual film was run through a log transformation, a process that turns a lognormal distribution into a normal one. The transformed data were then checked for normality using several methods including the Shapiro–Francia (Shapiro and Francia, 1972) and Jarque–Bera (Jarque and Bera, 1987) tests. Redfern also uses several examples of graphical analysis including normal-quantile plots of the log-transformed data to argue against precise lognormality on a film-by-film basis. His article uses the combined metrics to conclude that most films tested (125 of 134 films) do not meet the stated criteria of lognormality.

2.1 Interpretations

Redfern also claims that characterizations of shot lengths as lognormal (Salt, 2006, 2011) are methodologically unjustified. In the discussion, Redfern argues that shot lengths of Hollywood films do not precisely follow a lognormal distribution, and thus, using parametric statistics may be either too difficult to interpret or completely inappropriate to apply to shot-length data. Redfern also attempts to refute the usefulness of the mean or median as an indicator of film style, proposing that interquartile range is a more appropriate metric.

3 Response

3.1 Interpretation

The major shortcoming within Redfern's article lies within the interpretation of the normality test results.

While it is absolutely true that shot lengths do not follow a strict lognormal distribution, refuting all utility of a lognormal model is an overstatement of the results. Statistician George E. P. Box (Box and Draper, 1987, p. 424) argued that 'Essentially, all models are wrong, but some are useful'. In a broad sense, showing that shot-length distributions do not 'perfectly' follow a lognormal distribution is unsurprising and ultimately uninteresting.

Almost all data collected from the world can be shown to vary from a perfect distribution given large-enough samples. A demographic example is the height of an average male, widely described and accepted as following a normal distribution. According to the United States Department of Health and Human Services *et al.* (2006) the average height of a male in the USA is 168 cm (5' 6") and varies with a standard deviation of 22 cm (8.5"). If the real world 'perfectly followed' the expected normal distribution, we should be able to find roughly 160 men that are taller than 272 cm (8' 11"), the height of the tallest man ever recorded (Guinness World Records, 2009). Given that the NBA has not found and recruited these extraordinary individuals (they might stick out in a crowd), we can conclude that the tall men do not exist and our model is not perfect.

Showing a model has flaws is easy. Deciding whether the flawed model still explains enough data to be useful is a more complex issue.

3.2 Graphs and plots

The first check when fitting a distribution is to actually look at the graph and to visually inspect the distribution, one of the methods advocated in the method of Exploratory Data Analysis cited within Redfern's article (Tukey, 1977). It is a fairly commonplace practice to present a distribution and fit to let readers inspect goodness of fit. While this does not sound like a particularly compelling form of evidence, showing a distribution and fit line is a type of evidence that your model is in the ballpark.

Redfern characterizes the chapter as 'lacking in methodological detail and statistical rigour' because the only evidence is a single graph showing a shot-length distribution with a lognormal fit superimposed over it. If the chapter, at any point, made a strong claim that the shot lengths for each film perfectly follows a lognormal distribution, the critique might have been justified. The chapter characterizes shot-length distributions as highly skewed and approximately lognormal. The single graph shows the distribution, to punctuate the point made in the figure's caption that 'Because the distribution has a heavy positive skew, median shot length can be

seen as a more accurate description than ASL for shots in a film' and not to prove lognormality.

Redfern's criticism implies that a single graph is simply not enough evidence, stating that 'Naturally, we would not expect to find a paper presenting charts 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'. The supplemental materials of this response include a series of 402 graphs displaying the quality of a lognormal fit for the 134 films, using the visualizations advocated by Redfern including a non-transformed data set with a lognormal fit, a log-transformed data set with a normal fit, and a Q-Q plot on log-transformed data.

Using these graphical methods, it is easy to pick out an extremely poor fit. Figure 1 shows that 'In Pursuit to Algiers' (1945) is a film that most certainly does not fit a lognormal distribution. All three figures show the same shot-length data from three different viewpoints. Figure 1a shows the shot lengths on a standard linear scale and has a lognormal fit applied. Redfern argues that this format is difficult to interpret given the highly skewed distribution, and advocates that fitting a normal curve to log-transformed data is easier to assess, so this visualization is shown in Fig. 1b. Figure 1c assesses the quality of the normal fit to log-transformed data by plotting how the distribution of the actual data set differs (dark '+' signs) from what would be expected if the data perfectly fit the distribution (straight gray line). By using all three plots, we can see 'how' 'In Pursuit to Algiers' deviates from an ideal distribution, including more long shots and fewer of the shortest shots than a perfect distribution would suggest.

Visually picking out which films do and do not fit Redfern's criteria can be difficult: Fig. 2 shows 'Barry Lyndon' (1975), which is classified as lognormal, and Fig. 3 shows 'Three Days of the Condor' (1975), which is classified as not lognormal. The difference in the Q-Q plot of 'Three Days of the Condor' appears to deviate significantly from the expected values of a lognormal distribution at the smallest shot lengths. This means that the 'Three Days of the Condor' does not have as many ultra-short shots as would be typically expected by a lognormal distribution.

How prevalent are the shortest shots in our database? In a sense, they are fundamentally limited by the fact that film is composed of many still images presented quickly enough so the human visual system perceives fluid motion. In our database, films have been converted from DVD to a fixed rate, twenty-four frames per second digital video file. This means that each frame is presented for roughly 41.666... ms, providing a hard lower limit for shot lengths. Humans, however, will not be able to watch a sequence progressing that quickly for long (Cutting *et al.*, 2011) so there must be some

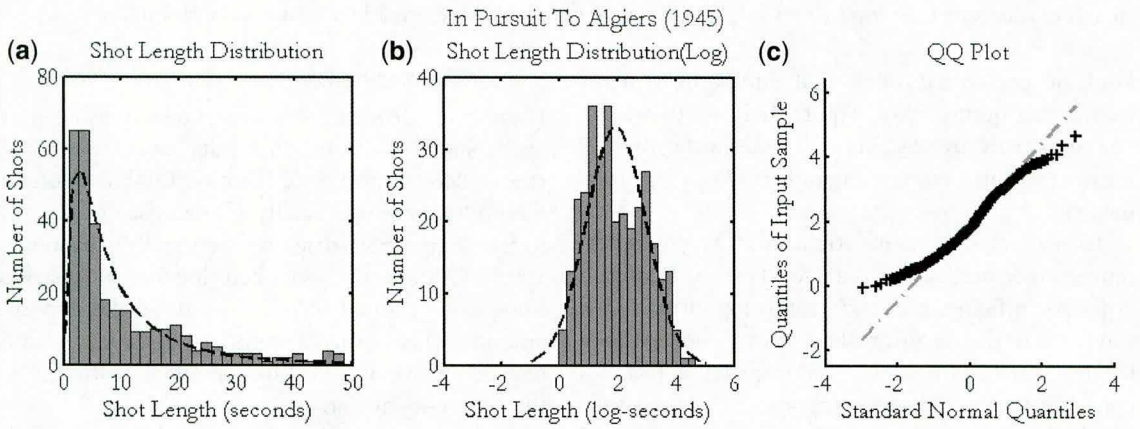

Fig. 1 'In Pursuit to Algiers' is a poor fit for a lognormal distribution. (**a** and **b**) show grouped values differing from the idealized distribution (dashed line). (**c**) The Q-Q plot reveals consistent deviation from the expected value line

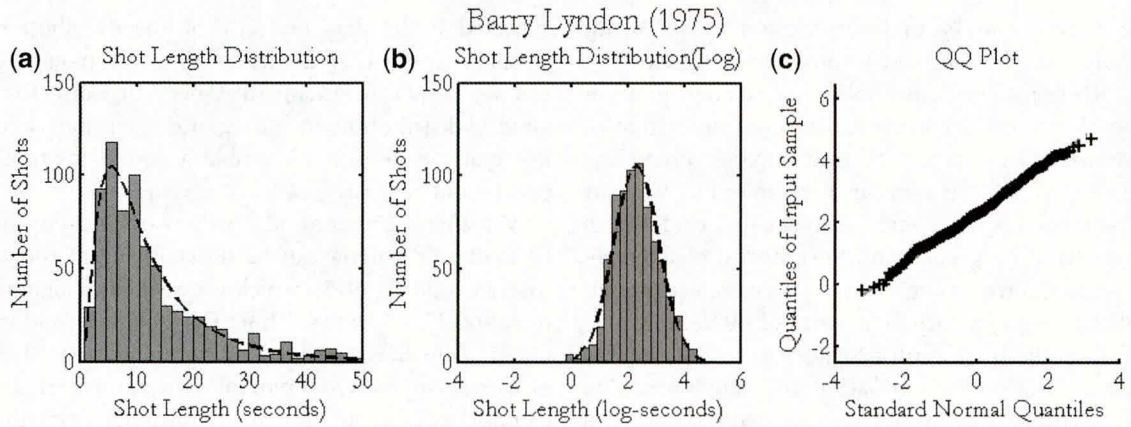

Fig. 2 While clearly not a perfect fit, 'Barry Lyndon' satisfies Redfern's criteria of lognormality

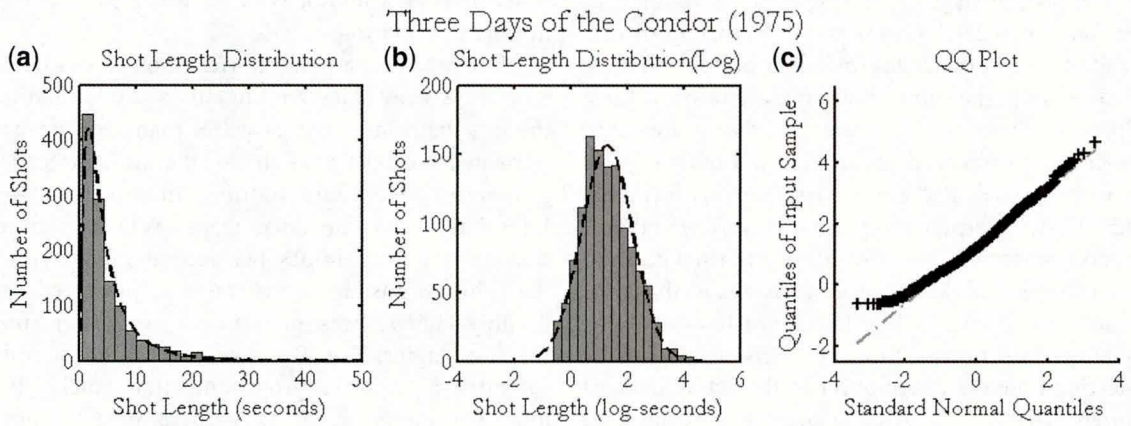

Fig. 3 'Three Days of the Condor' appears to be a good fit for a lognormal distribution; however, investigation of (**b** and **c**) shows that the lowest shot lengths are not as prevalent as expected in a lognormal distribution

kind of perceptual floor that limits how many frames can be in a shot. Figure 4 shows that in all 134 films there are few shots that are shorter than 12 frames (500 ms), comprising only 0.18% of the total data set.

How much of an effect could this tiny, yet underrepresented, part of the data set have on Redfern's criteria of whether a distribution is lognormal? One way to test this is to create a distribution of shot lengths that has the same basic statistics as the shot data, but that we 'know' is lognormal. We can then tweak this data, only removing the shots that are shorter than 12 frames long to mimic what we find in the database.

3.2.1 *The one-percenters*

Lognormal distributions were created using averaged statistics from the data set ($\mu = 4.5255$, $\sigma = 0.9046$, number of shots = 1,085). Redfern's methodology works exactly as advertised on the perfectly lognormal data set, correctly diagnosing 100,000 randomly generated lognormal distributions as lognormal 95% of the time, the expected amount when using a cutoff of $P < 0.05$. Clearly, having an average of 1,085 shots is sufficient for the Jarque–Bera test.

Redfern's methodology works differently if the ultrashort shots (those <500 ms, comprising ~0.9% of the total shots) are removed from the

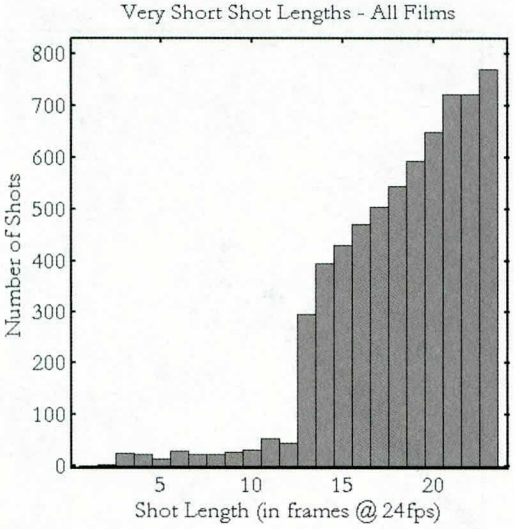

Fig. 4 The number of ultrashort shots for all 134 films, grouped together by length in frames. Films are much more likely to have shots that are longer than 12 frames (500 ms), possibly driven owing to perceptual demands for a minimum shot length

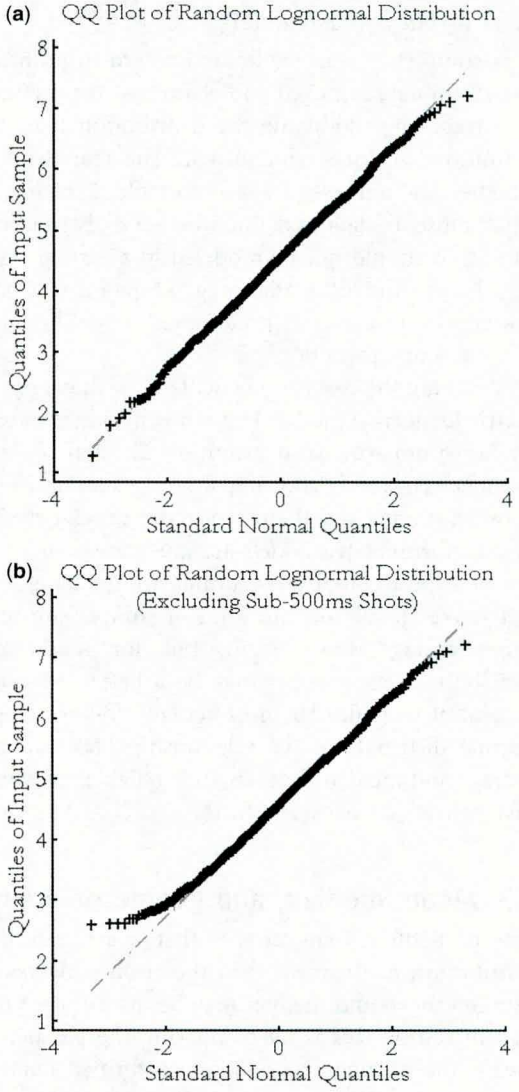

Fig. 5 Q-Q plots from a single simulated film generated from average database statistics. (**a**) shows a perfect lognormal distribution, while (**b**) shows what happens to the graph when shots shorter than 12 frames (500 ms) are removed. It is worth noting the similarity between the distribution of 'Three Days of the Condor' (Fig. 3b) and Fig. 5b

randomly generated distribution. Removing the small sliver of ultrashort shots causes the method to flag the previously perfectly lognormal distribution as not lognormal ~67% of the time. As shown in Fig. 5b, a Q-Q plot shows that the deletion of the shortest 1% of shots causes the 'tail-up' characteristic observed in many films. Redfern's methodology appears to be sensitive to changes in ultrashort shot-length range that is rarely used by filmmakers and borderline imperceptible in comparison with average shots.

Another issue with Redfern's metric is that when faced with an absence of the shortest 1% of shots, it is increasingly more likely to report that the distribution is not lognormal, even if the underlying distribution is the same. If a 500 shot film is perfectly lognormal except for having no ultrashort shots of less than half a second, Redfern's method will conclude the distribution does not qualify as lognormal ~21% of the time (given 100,000 simulations). When the film is 1,000 shots, the rejection rate jumps to ~62%, and a film of 2,000 shots will be rejected ~97% of the time after eliminating the shots shorter than half a second (only 1.2% of the data set). This built-in biasing of the results is unsurprising: tests of normality are designed to be used to examine the characteristics of small random samples.

3.2.2 *What is reasonable?*

It is completely reasonable for Redfern to point at the 1% of data removed and claim that the method is correct—by modifying the distribution it is, by definition, no-longer lognormal. The real issue is whether or not we are comfortable allowing a small subset of data to define whether a distribution should or should not be modeled in a certain way. In this case, the influential range of data is so small it seems odd to reject a pretty good model because of small imperfections.

What are the costs and benefits of having a parametric lognormal model? In a sense, having the ability to summarize a distribution in two simple numbers is useful, allowing for easy comparisons between groups and the use of more general statistics. Unfortunately, models are not perfect and can not be expected to always summarize the data perfectly. We already do this kind of comparison utilizing average shot length, but for a skewed distribution, the median may be a better estimate of central tendency. If films actually follow a lognormal distribution, the relationship between the average and median shot length is reliable, allowing easy conversion back and forth.

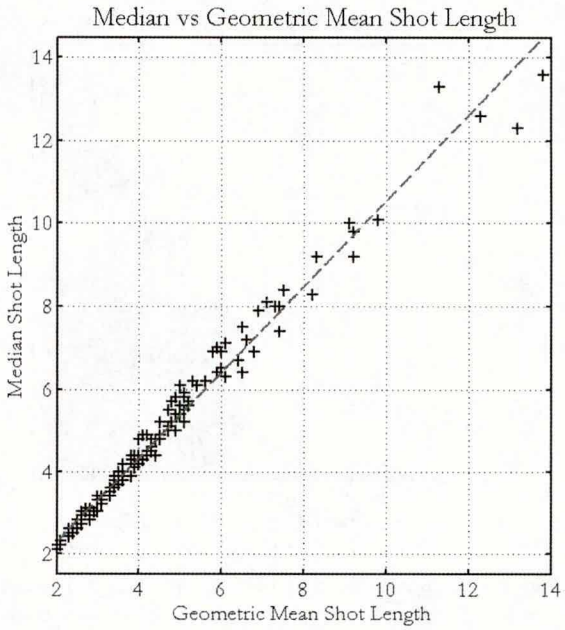

Fig. 6 Scatterplot showing the relationship between median and geometric mean. Redfern argues that the 'large discrepancies' between the two values are evidence against using the lognormal distribution; however, checking the relationship between these two measures shows a significant correlation. While it is true that the two metrics do not perfectly translate, they are clearly related

3.3 Mean, median, and geometric mean

One of Redfern's concerns is that if a lognormal distribution is assumed, then the clean conversion between mean and median may be inaccurate. One way of testing this is by comparing the geometric mean (the exponential of log-transformed curve's mean, which is to say, the middle peak of the log-transformed data) and the median of the original data. If the distribution is perfectly lognormal, these numbers 'should be equal'.

After comparing the geometric mean with the median values for each film, Redfern declares that 'These ratios are reliable evidence (by "modus tollens") that the data are not lognormally distributed and where we observe large discrepancies we can be confident the assumed model is not appropriate'. A paragraph later, Redfern admits that 'The problem of interpreting these results therefore becomes one of deciding what constitutes a "large discrepancy" between estimates'.

When looking at the median versus geometric mean values for the 134 films, it become immediately apparent that while not perfectly aligned, the large discrepancies described by Redfern are nowhere to be seen. The geometric mean is, on average, ~0.38 s longer than the median. As seen in Fig. 6, the two measures are 'very very' significantly correlated ($R^2 = 0.9799$, $P < 0.05 \times 10^{-112}$). This shows us again that while the majority of shot-length distributions are not perfectly lognormal, they are 'close' and outstandingly 'reliable'.

One way to see the big picture of shot-length distributions is to put all 134 films together onto one graph. Because different films tend to be cut at different rates, it is necessary to normalize them to a single value. Each of the 134 films was divided by its average shot length so that every film's transformed average shot length was one in an attempt to correct

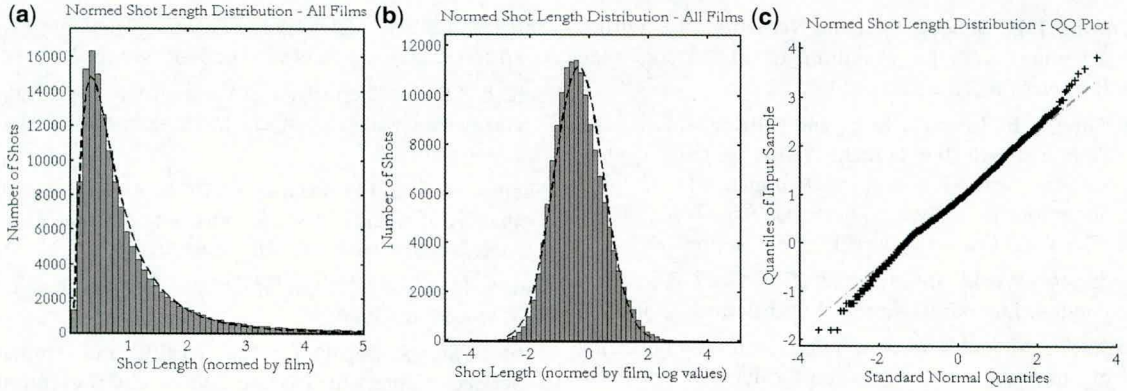

Fig. 7 Distribution graphs showing every shot (normed by each film's average shot length) for each of the 134 films. (**b**) especially highlights that the distribution is slightly positively skewed ($\gamma_3 = 0.31$ for log-transformed shots) than a perfect lognormal distribution; however, the resemblance is uncanny

for decreasing shot length over the years. All adjusted shots were concatenated into a single list and graphed in Fig. 7. The grand distribution is slightly more skewed than lognormal, explaining why the geometric mean was slightly larger than the median. Despite not being perfect, it is awfully hard not to conclude that the fit is, at the very least, a 'pretty good' approximation of the data.

4 Conclusion

Redfern's article attempts to refute the claim that shot-length distributions exhibit a perfectly lognormal distribution. Curiously, this claim has not been made by myself or Barry Salt. Redfern ultimately overextends his analysis, concluding that because shot-length distributions are not perfectly lognormal, the distribution is 'inappropriate'. Ultimately, the logic and conclusion of Redfern's article break down, but the initial concerns still have value. The question is not 'whether' shot distributions are definitely lognormal, it is 'how close' to lognormal they are. Judging by the data on the table, lognormal is not perfect but it is a pretty good fit.

The more interesting question is—why would shot-length distributions exhibit any structure (lognormal or not) in the first place? Arguing model fits does not get us any closer to understanding 'why' film is built the way it is.

As a cognitive scientist, I am biased to believe that the brain has impacted the structure of film, and that in examining the cinematic record we have the unique opportunity to find evidence of a young art form accommodating to our perceptual and cognitive architecture. We far from understand how our brains see the world, and film has many secrets yet to reveal.

Even if shot structure is not perfectly lognormal, the fact that a similar distribution shows up again and again hints at some deeper mechanism. Perhaps 'events' in the world come at us in a particular structure, and film is emulating the structure from the world. Perhaps our brains and eyes work in such a way that something close to lognormal makes sense. Maybe it is just pretty. Regardless, it is worthwhile to keep investigating whatever patterns we manage to dig up from the history of film.

References

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-building and Response Surfaces*. New York: Wiley.

Cutting, J. E., DeLong, J. E., and Brunick, K. L. (2011). Visual activity in Hollywood film: 1935 to 2005 and beyond. *Psychology of Aesthetics Creativity and the Arts*, **5**: 115–25.

Cutting, J. E., DeLong, J. E., and Nothelfer, C. E. (2010). Attention and the evolution of Hollywood film. *Psychological Science*, **21**: 440–7.

DeLong, J. E., Brunick, K. L., and Cutting, J. E. (2013). Film Through the Human Visual System: Finding Patterns and Limits. In Kaufman, J. C. and Simonton, D. K. (eds), *The Social Science of Cinema*. New York: Oxford University Press. In press.

Guinness World Records. (2009). *America's Tallest*. London, England: Guinness World Records, 2008.

Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, **55**(2): 163–72.

Redfern, N. (2015). The log-normal distribution is not an appropriate parametric model for shot length distributions of Hollywood films. *LLC. The Journal of Digital Scholarship in the Humanities*, **30**(1): 137–51.

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

Salt, B. (2011). The metrics in Cinemetrics, http://www.cinemetrics.lv/metrics_in_cinemetrics.php (accessed 24 January 2012).

Shapiro, S. S. and Francia, R. S. (1972). An approximate analysis-of-variance test for normality. *Journal of the American Statistical Association*, **67**: 215–6.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

United States Department of Health and Human Services; Centers for Disease Control and Prevention; National Center for Health Statistics (2006). *National Health and Nutrition Examination Survey (NHANES), 2005-2006. ICPSR25504-v5*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor] doi:10.3886/ICPSR25504.v5.