Film statistics: some observations

Mike Baxter, 16 Lady Bay Road, West Bridgford, Nottingham, NG2 5BJ, U.K.
(e-mail: michaelj.baxter@btconnect.com)

"The most popular type of quantitative film data to examine is average shot length (hereafter ASL), a metric of how long a shot is onscreen before transitioning to a new shot. ... Despite being the popular metric, ASL may be inappropriate because the distribution of shot lengths isnt a normal bell curve, but rather a highly skewed, log-normal 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 films shot length. A better estimate is a films Median Shot Length, a metric that shows the same decrease in shot length over time but provides a better estimate of shot length ..." [9].

1 Introduction

This paper is an edited version of something written earlier for my own benefit, to collect my thoughts on aspects of the cinemetric literature, arising from my recent exposure to the subject and fairly intensive reading around it. The original paper was written independently of the recent questions posed by Yuri Tsivian concerning the use of the median and ASL, but covers what I think are the main issues. To avoid a complete re-write I've retained some detail and discussion that is probably unnecessary for the current intended readership. An attempt has been made, for the most part, to stick to 'objective' statistical analysis. I've given freer rein to more 'subjective' views in the summary which addresses a little more directly some of the questions posed by Yuri. My lack of any background in film studies should be borne in mind when reading it.

There is a narrow focus on the choice and use of measures of location for shot length (SL) distributions, descriptive of some aspects of film style. The mode, median and (arithmetic) mean are all examples of measures of location. In cinemetrics the last of these is often called the average shot length (ASL), and it and the median are the measures most frequently advocated and used.

In the recent cinemetrics literature, as the opening quotation suggests, the relative merits of the ASL and median have been debated, the ASL holding historical sway and the median promoted in revisionist views. The main objections to the use of the ASL seem to be that it is highly sensitive to outliers in SL data, and that outliers are the rule rather than the exception. As *general statements*, claims about sensitivity to outliers and their prevalence are unsustainable; this is simply demonstrated empirically (Section 2).

Film shot length (SL) distributions are usually highly skewed. It is arguable – and so argued in Section 2.3 – that perceptions about the prevalence of serious outliers are based on misconceptions about what samples from such distributions should look like.

It should be axiomatic that for skewed distributions no single measure of location is adequate. One theme of the present paper is that it is better to use the median and ASL in tandem rather than treat them as 'competing' statistics from which a choice should be made. Section 3, which is mostly separate from the outlier issue, revolves about this. Another theme, counter to some suggestions in the literature and also explored empirically, is that the relationship between the ASL and median is often sufficiently strong that for comparing 'style' between films it often doesn't matter which is used; where it does, the merits of using either can be questionable. (To put this another way, it appears that you can sometimes get away with using a single measure of location, but it doesn't much matter which is used in such circumstances.)

It is interesting to ask why the merits about different summary statistics have become an issue at all. Section 4 examines this. Some of what has been written in the literature is based on assertions that don't always bear close scrutiny. The more evidence-based arguments are more interesting, but the evidence can sometimes be interpreted differently from what it is intended to show.

What follows is underpinned by the thinking that calculating summary statistics for a single SL distribution is pointless. Such activity is only valuable if used for comparative purposes. It often doesn't matter, for comparison, whether the median or ASL is used; where it does matter it is more useful to examine them jointly, and it is possible that neither may be suitable for summary and comparative purposes.

Some important topics are limited to brief discussion. Measures of location ought not to be used in isolation from each other, and should not be used in isolation from measures of scale (spread) – the temptation to discuss the latter in any detail has been resisted. Detailed discussion of 'robustness' is also limited, the median usually being advocated on the grounds that, compared to the ASL, it is a robust statistic. I don't regard it as self-evident that robustness – insensitivity to 'unusual' data – is 'a good thing'; sensitivity, of the kind manifested by the ASL, has its virtues as well.

2 Empirical results

The presentation is structured as follows.

1. For some films neither the ASL, median or any other simple summary statistic is an adequate descriptor of style (Section 2.1).

2. Eliminating these, there is often a reasonably good linear relationship between the ASL and median. Thus, it doesn't much matter which is used for comparative purposes (Section 2.2).

3. The suspicion exists that the prevalence of outliers has been exaggerated, because of misconceptions about what constitutes an outlier in data sampled from skewed distributions (Section 2.3).

4. Outliers often have no impact on ASL calculations that seriously affect any comparisons one might wish to effect with them. Given the previous point the implication is that the ASL and median will often lead to similar conclusions, even if outliers are present (Section 2.4).

Section 2.4 alone is enough to show that *generalisations* about the pervasive and deleterious effect of outliers on ASL calculations are, for practical purposes, unsustainable. It follows that arguments against the use of the ASL that are based on generalisations with this foundation are not, by themselves, sustainable. None of this is to deny that outliers *may* sometimes seriously affect ASL calculations, but it can then equally be argued that the lack of effect of such outliers on the median means that the latter is concealing interesting aspects of 'style' that the ASL is sensitive to.

For analytical purposes a subsample of 150 top-grossing Hollywood films, 10 from each of 15 years at five year intervals beginning in 1935, collected and analysed by James Cutting and colleagues ([3]-[7], [9]) is used. Of these, and following [15], 16 of the 150 are discarded because of data recording issues (e.g., zero values) making logarithmic transformation impossible. Thus the sample consists of 134 films – this is not, nor does it purport to be, a random sample from any well-defined target population, but this is not problematic for the uses to which it is put here.

2.1 Simple summaries aren't always useful

There are SL distributions for which no simple summary is suitable, occurring when they are 'lumpy' or not 'nicely smooth'. As well as obvious 'lumps' such as more than one mode, in an abuse of the term, 'lumpy' is used to describe distributions with shoulders, possibly even sloping, that detract from the aesthetic pleasure that can be derived from looking at some distributional shapes. This is best understood by looking at the kernel density estimates (KDEs) for selected SL distributions after a logarithmic transformation, as in Figures 1 and 2[1].

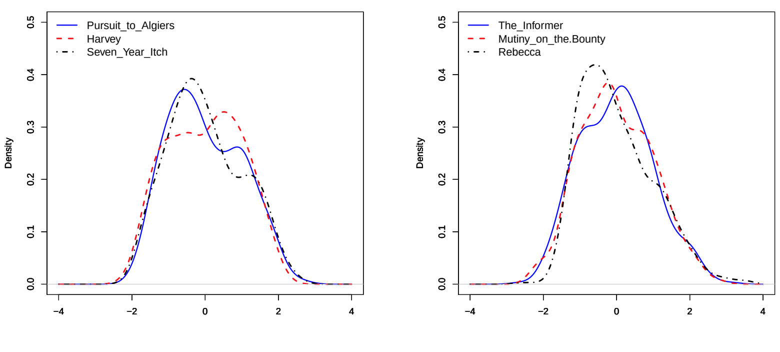

Figure 1: *KDEs for log-transformed SLs of selected films.*

It should be fairly obvious that neither the ASL nor median are suitable descriptors for the films in the left panel of Figure 1; I'd contend the same is true for most of the other KDEs in this and Figure 2. Some in the right of Figure 2

[1]For those unfamiliar with some of this, the KDEs can be thought of as histograms, with their unsightly blocky nature and corners smoothed out. There is a moderately extensive literature concerned with defining 'bumps', 'lumps'. 'smoothness' etc. in a much more mathematically rigorous fashion, and testing for their 'reality' on the basis of sampled data. Quite a lot of this literature is technically demanding, and much of it has had little application outside the papers in which the work is published, and then sometimes only for 'illustrative' examples.

are debatable; *East of Eden* (1955) and *Ocean's Eleven* (1960) are the later films illustrated, and the lump in the right tail for the latter is not especially prominent. Technical discussion follows; those prepared to accept the above can jump to the final paragraph of the section.

Logarithmic transformation is a convenience to make things easier to see. There is a monotonic relation between SLs and their log-transformed values, meaning that the rank order remains the same after log-transformation. Pattern on the untransformed scale may be easier to see after transformation. That is, if 'lumps' are visible on the log-transformed scale they exist in the untransformed data but may be less obvious. The data used to produce the KDEs have been standardised to zero mean and unit variance. That is, location and scale effects are removed so that comparisons and comment are based entirely on distributional shape.

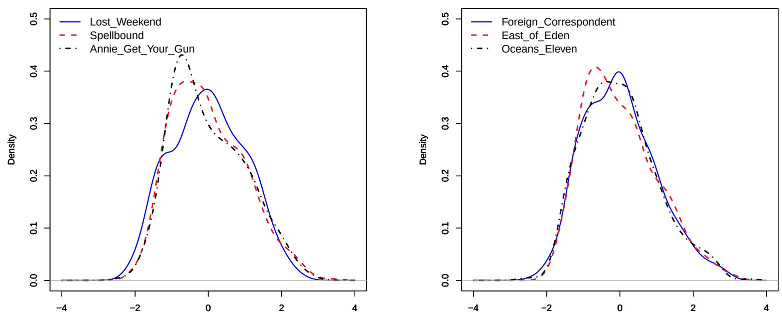

Figure 2: *KDEs for log-transformed SLs of selected films.*

If 'lumpiness' occurs, it is natural to think in terms of the data as a mixture of distributions. Assuming a two-component mixture, a descriptive summary could be obtained by separating out the two components and summarising them separately. No single descriptive statistic, be it the median or ASL, is appropriate. Note, also, that the focus is usually on what is going on in the middle of the distribution; KDEs aren't especially good at revealing the details of tail behaviour, for which probability plots are better.

How prevalent is the phenomenon under discussion? Judgements are subjective; mine is that 20-25% of films have 'lumpy' distributions, all 1980 or earlier, and most 1965 or earlier. Of nine films with fairly clear bimodal distributions all are 1960 or earlier. Taking the lower estimate, it is arguable that any discussion about the merits of single summary statistics is moot for about a fifth of the films.

2.2 What's all the fuss about?

Figure 3 shows plots of the median against the ASL, for films with 'non-lumpy' distributions, and for the 54% of films with the number of shots $n > 1000$. The (Pearson's) correlation r and Spearman's rank correlation are about 0.95 for both subsets. This suggests that the information from the ASL and median is broadly similar in the sense that, for any *comparison* we might wish to effect between films, it largely doesn't matter which of the median or ASL we use.

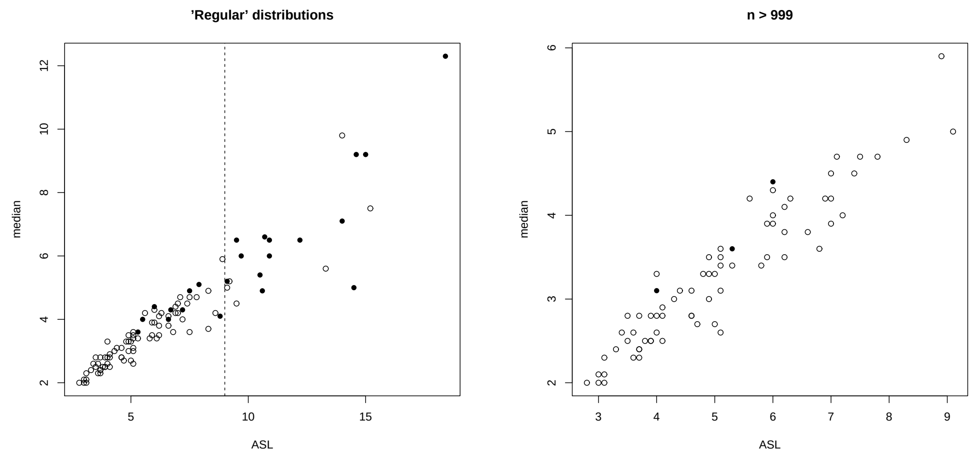

Figure 3: *Plots of median against ASL for all films with 'non-lumpy' distributions, and films with* $n \geq 1000$*. Solid circles are films earlier than 1960, the vertical line is for ASL = 9.*

Dates of the films are indicated, solid circles for films earlier than 1960 very few of which have $n \geq 1000$. The earlier films are those most deviant from linearity in the plot based on distributional shape, mostly where their ASLs are more than 9.

A comparison of the two plots suggests that in the right-hand plot, for $ASL < 9$, the films most deviant from linearity have $n < 1000$. Given the linearity observed it is worth pursuing the analysis of the films with $n > 1000$ a little further. The median/ASL ratio is normally distributed (Shapiro-Wilk test, p-value = 0.98), with a mean of 0.66 and standard deviation of 0.065. That is, for films with n of this kind of size, the idea that the median SL tends to be about two-thirds that of the ASL (obviously with variation about this) is sustainable.

The next two sections demonstrate that variation in the ASL induced by extreme values, whether classified as outliers or not, is usually not sufficient to require that the ASL be rejected in favour of the median for comparative purposes. Direct assessments of the relation between the median and ASL using other samples of films are examined in Section 4.

2.3 On the prevalence of outliers

Despite a copious literature on the subject (e.g., [2], [11]) a comfortably concise definition of *outlier* enabling everyone to reach the same decision about what is an outlier is elusive. Outliers have been defined as an observation 'that appears to deviate markedly from other members of the sample in which it occurs', [10]; an observation that appears 'inconsistent with the remainder of the data', ([2], p. 7); and elsewhere as an observation that causes 'surprise' in relation to the majority of the sample ([20], p. 119)).

'Surprise' may be relative to some presumed underlying model, such as a lognormal distribution, or because observations look unusual in relation to the body of the data. There is a suspicion that assertions about outliers in SL data are occasioned by surprise of the latter kind, and the suspicion also exists that it is frequently both understandable and unwarranted.

It is convenient to pursue this thought, initially, assuming that SLs can be treated as samples from a lognormal distribution. Figure 4 shows, to the left, SLs simulated from five known lognormal distributions, and to the right their logarithmic transformation.

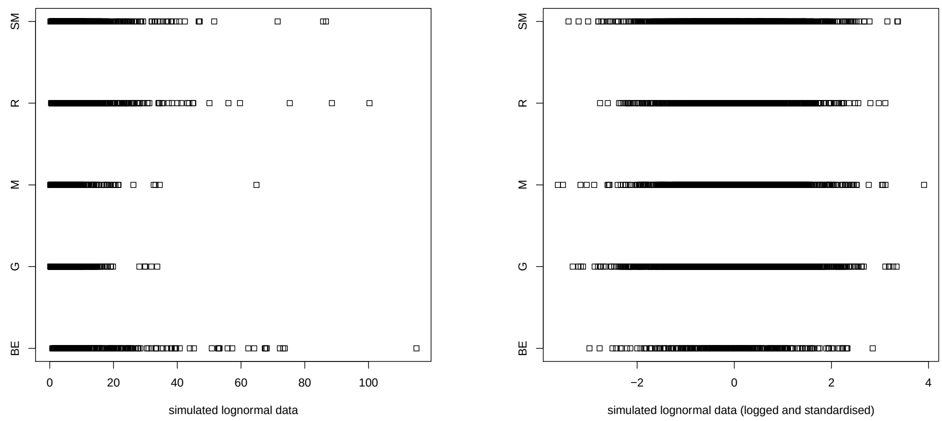

Figure 4: *Plots of simulated SLs for five 'films' before and after log-transformation. See the text for a detailed explanation.*

To add a certain realism to the exercise the sample sizes and parameters that define the distributions are based on five films, BE = *Brief Encounter*, G = *Goldeneye*, M = *Madagascar*, R = *Rebecca* and SM = *The Sound of Music* (there is no implication that the real films have lognormal distributions). If asked to say whether or not outliers are present for these films I suspect (and may be wrong) that many would answer 'yes' – most likely for two, or four, films and possibly all five. Any such assessement would be wrong. We know that the underlying populations are lognormal. We could be unlucky in sampling values that really are extreme, but the log-transformed data suggests not.

The lognormal distribution is smoothly varying; samples from it are not, and it is quite natural for some observations sampled from such a distribution to be at some distance from the main mass. In graphical representations such cases can appear isolated, giving rise to 'bumpy' tails in a histogram representation for example. The incorrect assessment that such cases are outliers would often be removed by looking at the data on a log-transformed scale. This suggestion is valid for underlying distributions qualitatively similar to the lognormal.

Rejection of the above argument does not invalidate the one to follow, that the *effects* of outliers where they exist have often been exaggerated. If the above argument is accepted it implies that not only have the effects of outliers been exaggerated, but also their *prevalence*.

2.4 On the effects of outliers

2.4.1 General considerations

Proceeding empirically, let SLs be ordered by size, from x_1 to x_n with x_n the most extreme of the n SLs. Let M_o be the mean of the O most extreme values. If A_O is the ASL calculated omitting O extreme values (whether outliers or not), and A = ASL uses all the data, the effect of omitting O extreme values with n large is

$$(\mathrm{A} - \mathrm{A}_O) \approx OM_o/n$$

where $\approx$ indicates an approximation is involved. There are 42 films in the sample post-dating 1980, 39 with $n > 1000$ and 30 with $n > 1200$. For the 92 films that are 1980 or earlier the comparable figures are 33 and 19.

The reader is invited to play around with this formula, real SL data for films with n in this sort of range, and either by omitting what they consider to be outliers or adding artificial but *realistic* outliers, see what they have to do to get a change in the ASL they consider to be of substantive significance. If, as I do, you conclude it is quite difficult to realistically induce serious changes, you are also left with the conclusion that for about half ($n > 1000$) or a third ($n > 1200$) of films the issue of the effect of outliers on ASL calculations is largely irrelevant.

Most of the films with 'lumpy' distributions, about 30, have $n < 1000$. Adding these to the numbers used above, the conclusion is that for 60-80% of films, the higher figure being perfectly reasonable, doubts about the value of the ASL arising from concerns about the effect of outliers are either moot or irrelevant. The figures being cited obviously depend on the sample being used but, if the arguments are accepted, with some generality it can be concluded that for most films from, say, the 1970s on, which are not 'lumpy' and have large $n > 1000$, the effect of outliers on the ASL is not a problem.

2.4.2 Single extreme values

The main objection to the above line of reasoning I can think of is that, for any single film, it is possible to construct examples that demonstrate outliers are a problem; what do the films themselves have to say about this? Using exact calculations, Figure 5 shows the change in ASL occurring on omission of the *single* most extreme value plotted against n and ASL for each film.

The dashed lines, at a change of 0.12 deci-seconds in both plots, $n = 900$ and $ASL = 9$ are arbitrary but designed to emphasise that for the majority of films very small changes occur for ASLs less than 9 and/or $n > 900$. The subsets so defined are not identical, but there is a lot of overlap and they tend to be films later than 1965. There is, incidentally, no implication that the most extreme value is an outlier.

2.4.3 Multiple extreme values

The demonstration in the previous section is a sufficient refutation of any claim, aspiring to generality, that the ASL is an unsuitable summary statistic because of the potential effect of single outliers; lack of robustness is a theoretical concern of

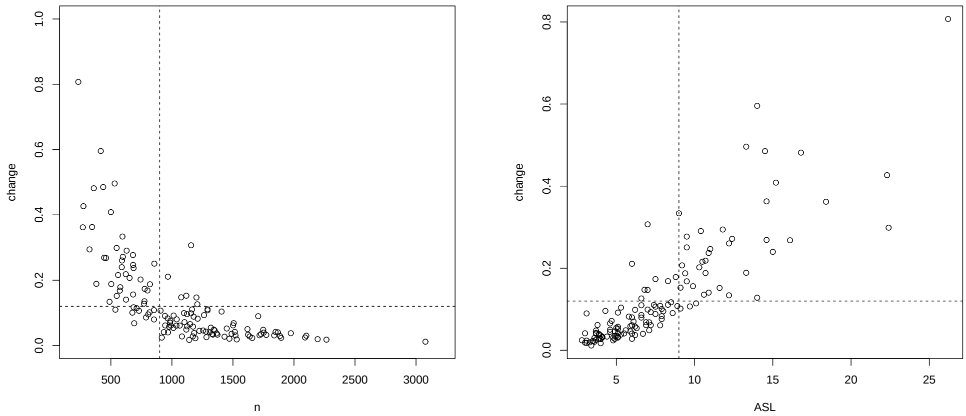

Figure 5: *Plots of the change in ASL on omitting the most extreme SL, against n and ASL. Horizontal lines are at 0.12s, vertical lines at* $n = 900$ *and ASL = 9.*

little relevance[2]. The wording is careful; there have been objections to the use of the ASL not solely dependent on the effect of outliers. This is dealt with in Section 3. There is no intention of denying that outliers *may* have an adverse effect on ASL calculations and inferences drawn from them – all that is claimed is that some of the literature grossly exaggerates the extent of the problem.

Turning to the effects of multiple extreme values, Figure 6 is similar to Figure 5, except that the change is based on the omission of the five extremes. Five is an arbitrary choice; individual examination of each film suggests that most have fewer than five outliers, if any.

The guidelines are again arbitrary. Of the 21 films above or cut by the 0.8s change boundary, 20 have $n < 650$ mostly with ASL > 12. Of these 20 all but *Catch 22* (1970) are 1960 or earlier. There are 11 films previously categorised as 'lumpy' where there should not be an issue about using the ASL since neither it nor the median are suitable summary statistics. The 11 include *The Seven Year Itch* illustrated in Figure 1 and for which the change is greatest. Other 'lumpy' films with the larger changes illustrated in Figures 1 and 2 are *Annie Get Your Gun*, *Harvey*, *Pursuit to Algiers* and *Spellbound*.

Four of the remaining 10 films are acceptably lognormal [15]; these are *Brief Encounter*, *Detour*, *Exodus* and *The Great Dictator*. Given the lognormality, the effect of outliers on the ASL is not an issue.

This leaves five films (with ASL, change and percentage change in brackets) *Catch 22* (13.3, 1.7, 14%), *Inherit the Wind* (15.2, 1.6, 11%), *The Letter* (14.0, 1.5, 11%), *Mr.Roberts* (12.2, 1.2, 9%) and *The 39 Steps* (8.8, 0.8, 9%). Figure 7 shows plots for the SLs of each of these films, and those that are approximately lognormal. Of the possibly problematic films, to the left of the figure, three have what might be judged as clear outliers; for two, *Mr. Roberts* and *The 39 Steps* the changes in

[2]The theory, relating to 'breakdown' points, essentially says that you can make the ASL as large as you wish by adding an extra arbitrarily large SL to the film. Factor in realism, as in the exercise suggested in Section 2.4.1, and the lack of practical relevance of the theory in this context is apparent.

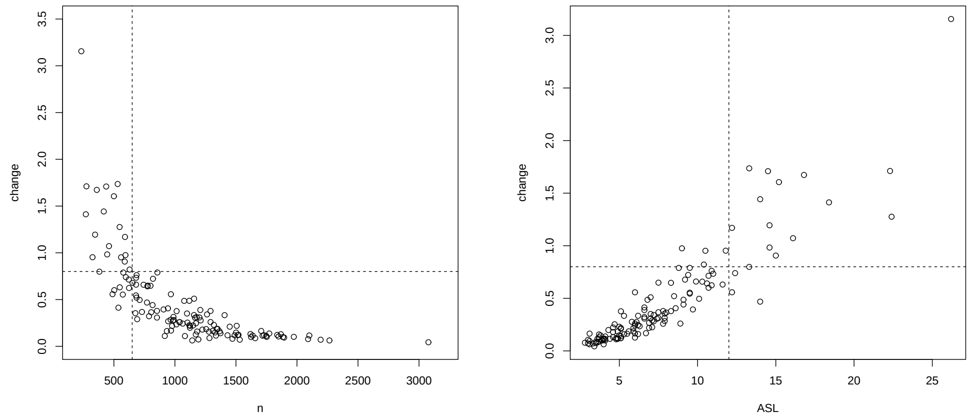

Figure 6: *Plots of the change in ASL on omitting the five most extreme SLs, against* n *and ASL. Horizontal lines are at 0.8s, vertical lines at* $n = 650$ *and ASL = 12.*

the ASL arise from the omission of SLs in the tail that are not obviously outliers.

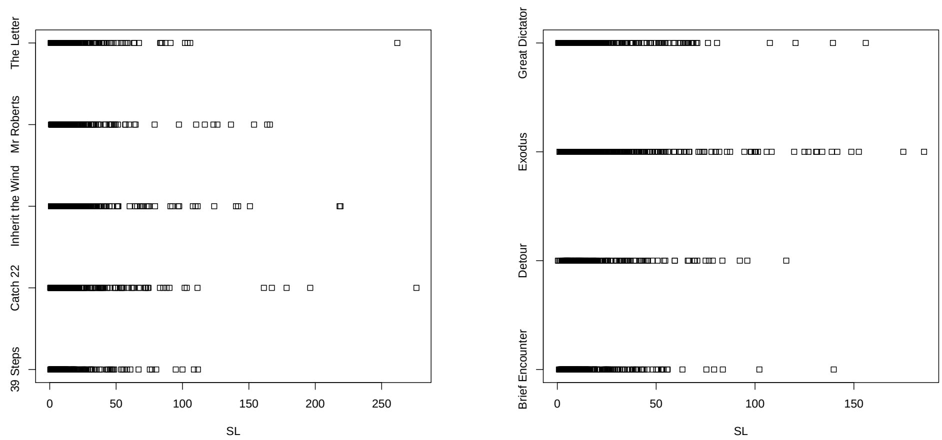

Figure 7: *To the left SLs for selected films whose ASLs are most affected by extremes; to the right ASLs for films with a lognormal distribution.*

Are the changes something to worry about? No generalisation is possible – it depends what the purpose of calculating ASLs is for. Figure 8 shows the effects of omission. The ASLs for each films are indicated by solid lines, their ASLs after omitting the extremes are indicated by the dashed lines to the left of each solid line. They are shown against the background of the ASLs for all films, truncated at each end so that the resulting change of scale makes it easier to see the effect of removing extremes. Two graphs are used to avoid overlap in the ranges defined by the ASLs and their values after omission of the extremes.

Reading from left to right across the two plots the films are *The 39 Steps*, *Catch 22*, *Inherit the Wind*, *Mr. Roberts* and *The Letter*. For none of the films is the

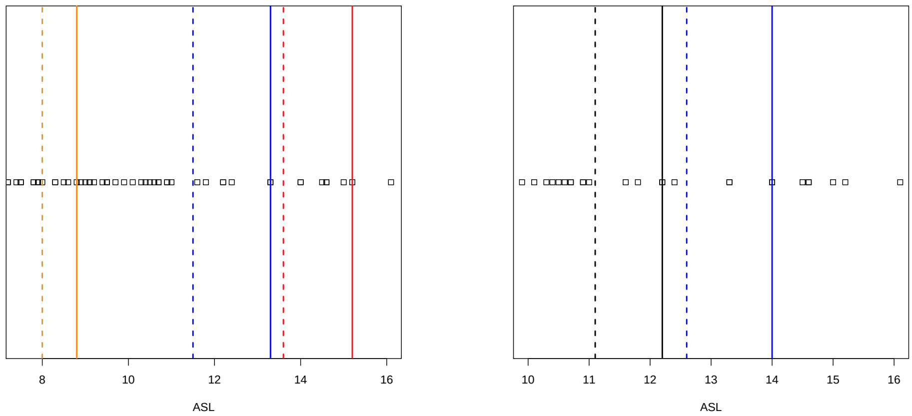

Figure 8: *Plots of ASLs with the ASL of five films before and after omitting the five most extreme SLs indicated. See the text for a detailed explanation.*

rank ordering dramatically changed relative to other films; about five places is the maximum. The films with the larger absolute changes occur at the larger and more thinly populated end of the ASL scale so the larger changes in the ASL have little impact on relative rank. Assuming that ASLs are calculated to effect comparisons with other individual films or bodies of films it is likely that the changes illustrated would have little influence on comparisons made unless very fine distinctions were sought.

3 Skewness

3.1 Lognormal distributions

The fact of skewness, independently of the existence of outliers, is sometimes cited as a reason for preferring the median to the ASL. This issue is dealt with here. It is convenient, in the first instance, to assume an underlying lognormal distribution, details of which are given in the Appendix.

Both the lognormal and Normal distributions are *completely* specified by two parameters, (μ, σ). For the Normal distribution the mean, median and mode of the distribution equal μ, all being measures of location. The parameter σ is a measure of scale or spread with the helpful interpretation that about 95% of the distribution lies within 2σ of μ.

It is sometimes useful, in exploring properties of SL data, to work with their log-transforms, for example, in examining whether SL distributions are indeed lognormal [15]. Summary statistics for the untransformed data have more direct meaning in terms of the properties of the films. Using the notation $(\mu_L, \Omega_L, \Phi_L)$ for the mean, median and mode of the lognormal distribtion and σ_L for the standard deviation

$$\mu_L = \exp(\mu + \sigma^2/2) \quad \Omega_L = \exp{(\mu)} \quad \Psi_L = \exp(\mu - \sigma^2)$$

with

$$\Psi_L < \Omega_L < \mu_L$$

and

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

Here, $\exp(\bullet)$ is the exponential function, and we can also write $\mu = \ln \Omega_L$. Note that μ_L is the ASL.

It follows from the dependence of the lognormal on two parameters that no single statistic can provide an adequate summary. The median, as a function only of μ, can be regarded as a 'pure' measure of location, but is totally uninformative about scale differences; the ASL conflates location and scale effects. The two jointly, (μ_L, Ω_L), provide information on both location and scale differences when used comparatively, and their use in combination, rather than regarding them as being in opposition, makes much more sense. Other pairs of statistics that involve μ and σ might equally be used.

In everyday language the mode is often referred to as the most 'common' or even 'popular' value, the median as the 'middle' or 'central' value (of ordered data). That the mean is the 'centre of gravity' of a set of data is less common parlance and less 'approachable' in terms of immediate understanding. There is the suspicion (Section 4) that the median is sometimes preferred to the ASL (the mode being strangely neglected) for precisely this reason; pandering to populist linguistic sentiment is not a good basis for mediating between the claims of different summary statistics. This, with empirical illustrations, is pursued in Section 4.

3.2 Skewed distribution, not lognormal

The point has already been made in 2.1 that if a data set has more than one mode or is at all 'lumpy' most of the usual measures of location and scale are not useful. Such distributions should be readily detected in preliminary data analysis and are not considered further in this section.

This leaves us with observed distributions having a single peak, skewed with a long(ish) tail to the right for which the assumption of lognormality is deemed invalid. For single-peaked nearly symmetric data sets that are not normal, methodology only strictly valid for the normal distribution nevertheless often works quite well for practical purposes. In the same way, if you think (even wrongly) that a distribution looks as if it might be close to lognormal, using methods only strictly valid for the lognormal may do little harm.

More explicitly, using the notation $(\widehat{\mu_L}, \widehat{\Omega_L}, \widehat{\sigma_L})$ for the calculated mean, median and standard deviation of a data set, without regarding them as estimates of parameters from some underlying model, summaries based on pairs of these – as suggested for the lognormal – may be almost as informative as one would wish.

If this is not acceptable there is a plethora of descriptive statistics to choose from. If you want to say anything useful about a skewed distribution in terms of measures of location you need more than one statistic. Where you need to start thinking beyond the lognormal as an approximation it is possible that no simple summary is suitable, and direct examination of the distributions concerned is needed. This is pursued in Section 4.

4 Arguments for the median

4.1 'Arguments' from assertion

The superiority of the median to the ASL has been asserted, or unquestionably accepted, often enough for it to be worth reviewing the arguments used to support the case.

It is sensibly noted in [9] that in looking at changes in SLs over time it doesn't matter whether you use the ASL or median. This is, nevertheless, accompanied by an expression of a preference for the latter. The assertion, aspiring to generality, that 'a small number of *remarkably* long shots inflate the mean' (my emphasis) is not true (Sections 2.4, 2.3). The statement 'that the large majority of shots in a film are actually below average, leading to systematic over-estimation of individual films shot length' is about a property of the lognormal distribution (which they assume) that more SLs are below the ASL than above, that is true by definition. There is the clear implication that a statistic with equal numbers of SLs either side of it is to be preferred. This, of course, is the median.

This 'argument' is saying that the ASL is not the median, with the implication that the median is to be preferred because it is the median. If not a tautology, it is an expression of what is being regarded as a self-evident truth. I dwell on this because it is not untypical of what can be found in the cinemetric literature. It is not an argument for preferring the median to the ASL.

Among papers that have been cited (in [16]) in support of using the median in preference to the ASL are Adams *et al.* [1], Kang [12], Schaefer and Martinez [18] and Vasconcelos and Lippman [19]. Reading these doesn't inspire confidence in the support they provide. They are largely unsullied by references to relevant statistical publications. The one exception, ([19], p.15), stating that it 'is well known in the statistics literature that the sample mean is very sensitive to the presence of outliers in the data. More robust estimates can be achieved by replacing the sample mean by the sample median' references [17] which is marginally relevant, if at all[3].

The paper uses the word 'median' precisely once, a feat emulated in Kang ([12], p.245) whose contribution 'We compute each shot's length and compared it with the median shot length because it shows a better estimate of the average shot length in the presence of outliers' is distinguished, though not uniquely, by sloppy and undefined use of the term 'average'. If you wanted to impose sense, of a sort, on the sentence 'we use the median because it's the median' might be it.

Adams *et al.* ([1], p.72) regrettably also confuse terminology by stating that the median 'proves a better estimate of the average shot length in the presence of outliers'. They make considered use of the median in devising what they call a 'tempo function', and its use is merited in context, but doesn't constitute general support for preferring the median to the ASL.

The assertion that the 'median shot length figures ... represent better indicators of shot length than the means, because the means are *inordinately* influenced by a few outlier values from the longest shots' ([18], p.335); my emphasis again)

[3]The text referred to is about outliers in the context of linear regression; none of the 25 or so real data sets chosen to illustrate problems arising from outliers have $n > 50$, the vast majority having $n < 30$. The relevance to SL distributions with n in the 100s or 1000s is not transparent.

would be mitigated if 'may be' replaced 'are'; as it stands it's an unthinking assertion of an unsustainable generalisation. The paper has the merit of providing enough information that allows the claims about effects of outliers to be seen to be unjustified.

The medians and ASLs are reported for 16 analyses relating to four editing variables descriptive of newcasts from four separate years, the interest being in hypothesised changes that have taken place over time. The patterns of change as revealed by the rank order of the median or ASL are virtually identical, minor differences arising from differences in absolute values of 1 deci-second. For all years apart from 1969 (the earliest) the ASLs are less than 8s, and within editing variables the median/ASL ratio is stable, so there is no indication of outliers having an effect. The figures for 1989 don't disrupt the general pattern since they are typically larger; for three variables the median/ASL ratio is smaller than for other years. There is not enough detail to see why but the phenomenon, revealed by a comparison of the ASLs and medians, is arguably more interesting that anything suggested by either statistic separately. Evidence for the ASL being a misleading indicator of 'style' is underwhelming.

4.2 Arguments from empirical data analysis

The most consistent advocate for using the median (and associated measures of scale) in preference to the ASL that I have read has been Redfern (e.g., [13]-[16]). These papers are of interest in providing what is intended as empirical support for the claims made for the superiority of the median to the ASL. They have the merit, also, of providing sufficient methodological detail, and data access, to allow independent assessment of the analyses. This is pursued below, with a view to showing that other interpretations of the data are possible. These are by no means unfavourable to the ASL, though the main message is that the joint use of the ASL and median together is more fruitful than the use of either separately.

4.2.1 Example 1 – Early Hitchcock

The prelude to the main business of [13] involves two Hitchcock films, *Easy Virtue*, a late silent, and *The Skin Game* an early sound film. The two pages or so on this culminate in an informative comparison between the films, using cumulative frequency diagrams (CFDs) on a logarithmic scale, that tells you everything about their differences you might reasonably wish to know.

The use of CFDs (or density estimates[4] as in Figure 9, also on a logarithmic scale) shows the obvious differences between the films. The median and ASLs, on the untransformed scales are (5, 6.7) for *Easy Virtue* and (5, 17.4) for *The Skin Game*. Several features are obvious from these statistics and Figure 9. That the distributions have a similar 'central tendency' is evident, as is the fact that SLs for the *The Skin Game* are much more dispersed, spreading to either side of those for

[4] In previous figures KDEs have been used; here local polynomial density estimates have been used, emulating the code on p.132 of [20]. The ideas are the same but local polynomials model tail behaviour better, which was not previously an issue.

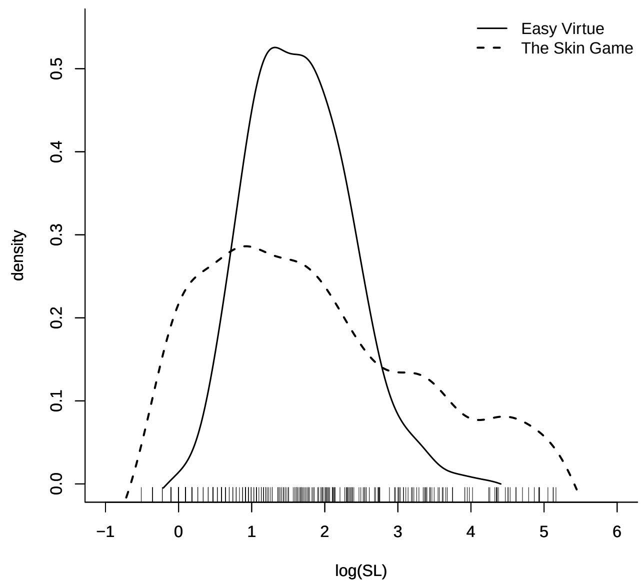

Figure 9: *Density estimates for the logarithms of SLs for* Easy Virtue *and* The Skin Game*. The 'rug' of data points shown is for the latter film.*

Easy Virtue. The 'lumpy' appearance of the right-tail of *The Skin Game* is also evident; the data points shown in the 'rug' do not highlight any outliers.

Graphical comparison obviates the need for agonising about which of the median or ASL is to be preferred. In [13] the example is used to suggest the ASL is not 'robust', is greatly affected by skew and outlying points, and that the median is 'much more reliable'. That the identity of the medians indicates the similarity of the films in terms of *one* aspect of style, centrality of the SLs, that the ASL doesn't is undeniable. It is, though, equally undeniable that the ASL alerts one to differences in another aspect of style, dispersion, that the medians completely miss.

Neither statistic is satisfactory by itself, and prompts the use of graphical comparison if this has not been done in the first instance. Neither the median or ASL should be used in isolation, and in [13] the inter-quartile range (IQR) and range are used to bring out aspects of dispersion that the median alone cannot address. It would be possible to devise comparable ways of enhancing interpretation of the ASL; in fact the lumpy nature of the distribution for *The Skin Game* suggests that neither the median nor the ASL is a particularly appropriate descriptor of style.

In [13] it is stated that the ASL for *The Skin Game* is about 2.6 times that of *Easy Virtue* and 'we might conclude from this that shot lengths in *The Skin Game* are typically longer than those of *Easy Virtue*' and this is 'not a sound inference'. We might and we shouldn't; you can't condemn a statistic just because it *might* be wrongly interpreted. It doesn't follow that 'the mean shot length is, therefore a poor indicator of film style'.

For this example – it is not generally true – the ASL and median appear to tell different stories, and *both* are failing in isolation to detect important aspects of style. This becomes apparent if they are viewed jointly, and prompts the use of graphical comparison, which then suggests that neither statistic is an appropriate descriptor for one of the films. That the ASLs might (incorrectly) be interpreted as implying that the 'typical' shot for one film is somewhat larger than for the other is not a valid argument against the ASL; you could equally (invalidly) argue that because the equality of the the medians might be interpreted by some as implying that distributions are similar it should not be used.

Essentially, it seems fruitless to argue that one statistic is better or 'superior' to the other; they do different jobs, can only provide partial information, and are much more informative used jointly, if distributional shape doesn't render both inappropriate.

4.2.2 Example 2 – Silent and early sound films

Here the relationship between the median and ASL for 20 silent films from 1921 to 1928 and 30 early sound films, 10 each from 1929 to 1931, is investigated. The data are from [13]. Figure 10 is a plot of the median against the mean for these data, labelled according to whether they are silent or sound.

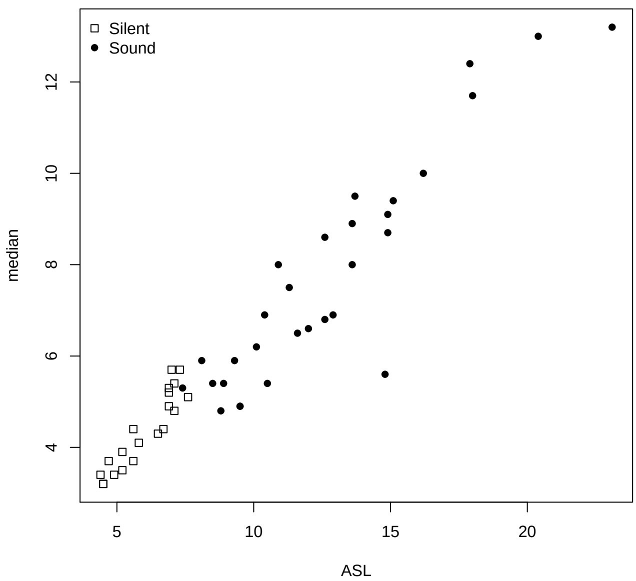

Figure 10: *A plot of median against ASL for 20 silent and 30 early sound films.*

With the exception of one sound film, *Applause*, the linear relationship between the median and ASL is good with the (Pearson's) correlation coefficient, r, and Spearman's rank correlation all about 0.93 for silent and sound films separately, omitting *Applause*. This suggests, following earlier arguments, that in any comparative study it would make little difference whether the median or ASL is used.

In [14] confidence intervals (CIs) are computed for the median for each film separately, and a count is then made of the number of films for which $c \times$ ASL lies in the confidence interval, using different values of c. For example, with $c = 0.6$, 4/20 films lie in the CIs, and 21/30 for sound films; for $c = 0.9$ the comparable numbers are 13/20 and 21/30. The approach doesn't allow for sampling variation in the ASL, and is specific to a limited number of values of c.

A regression through the origin of the median against ASL for the silent films suggests using $c = 0.73$ for which 18/20 films lie in the CIs; the other two are sufficiently close that if any allowance is made at all for the fact that 0.73 is an estimate, based on quantities (medians, ASLs) that are themselves estimates, it can be concluded that all 20/20 films have a sufficiently close relationship between the median and ASL for it not to matter which is used.

For sound films, omitting *Applause*, $c = 0.62$ is suggested. This leads to 23/29 films lying in the CIs and three (possibly more) lying outside but close enough to the limits that any attempt to allow for sampling variation would admit them. Some of the remainder have sufficiently 'lumpy' distributions to cast doubt on the wisdom of using either the median or ASL.

The overall picture is that it doesn't much matter whether the median or ASL is used as a comparative basis for the analysis of 'style'. This is at variance with the conclusions in [14]; exactly the same data and essentially the same methodology has been used, with the important difference that the data has been allowed to 'speak for itself' in the choice of c, and sampling variation in the ASL is taken into consideration (admittedly rather informally).

5 Discussion

The applied statistics literature is replete with warnings about the *potential* effect that outliers can have on data analysis – along the lines that the sample mean can be 'completely upset by a single outlier' ([20], p. 120), whereas the median is not. None of this can be disputed theoretically; whether or not outliers have much of an *actual* effect – even if correctly identified as such – can be checked.

The reading of the cinemetric literature presented here, supported by analysis of data, and in some cases the methodology, that has been used by others, leads to the following suggestions.

1. The *prevalence* of outliers has been exaggerated; probably because of confusion about what constitutes an outlier in a skewed distribution – that is, between outliers and what occurs naturally in the tail of a sampled skew distribution (Sections 2.3).

2. For a significant minority of data the distributions are sufficiently complex to militate against the naive use of either the median or ASL as summary statistics. (Sections 2.1).

3. It is a trivial observation that outliers will 'inflate' the ASL. Where the calculation of the ASL seems sensible, simple empirical analysis is enough to show that such inflation will often have little influence on any comparisons one might wish to effect using ASLs (Section 2.4).

4. Where not vitiated by distributional considerations, use of the median or ASL for comparative purposes will often lead to similar conclusions (Sections 2.2, 4.2.2).

5. Where the median and ASL do lead to different conclusions in comparative studies it is a lot more fruitful to ask why than insist that one is better than the other.

6. Empirical analyses, intended to support the median in favour of the ASL, are open to alternative interpretations (Section 4.2). In Section 4.2.1 arguments similar to those intended to support the median can be invoked in favour of the ASL; in truth it is much more useful to use them jointly.

   The analysis in [14] is, I think, intended to show there is no or little evidence of a consistent relationship between the median and ASL; and therefore the ASL is 'wrong'. A perhaps more 'nuanced' application of the methodology suggests these conclusions may themselves be 'wrong' (Section 4.2.2).

7. Regardless of the outlier 'problem', summarising a skewed distribution using a single measure of location doesn't, in general, make sense (Section 3.1). There is evidence that you can often get away with one statistic using SL data because, contrary to arguments in the literature, the median and ASL often tell the same story; where they do not, trusting either in isolation is not recommended.

Two points should perhaps be emphasised to avoid misunderstanding. One is that there is no intention of denying the possible effect outliers *may* have on ASL calculations and their interpretation; merely that the pervasiveness of deleterious effects arising from outliers is usually just assumed and is grossly exaggerated. The other is that nothing in the paper should be read as an argument favouring use of the ASL over the median. If they don't tell the same story, as an absolute minimum you need to look at both (or statistics carrying equivalent information) or, more likely, look at the entire SL distribution rather than summaries of it.

6 Summary

I am guessing that if the resources we now take for granted, computational and others, had been available forty years ago a debate about the relative merits of the median SL, ASL and the effect of outliers would be unnecessary. Both might have been routinely calculated and used jointly at the outset.

It is often useful to think in terms of models, even if you don't believe them; the lognormal distribution is one of the simplest that can be proposed for SL distributions; it depends on two parameters; and, except in special circumstances, at least two summary statistics are needed to characterise any body of data for which the lognormal is proposed as a model[5].

An obvious choice for a pair of summary statistics would be estimates of the parameters (μ, σ); the pairing (median, ASL), as argued above, is an alternative,

[5] Other distributions can be, and have been, proposed of similar mathematical complexity and the same arguments apply.

more directly interpretable in terms of what the SL distribution looks like. The median, in the context of a lognormal model, is a simple transformation of μ but has nothing to say about the scale or shape of the distibution, aspects of which are embodied in the ASL.

I assume that, in reality, anyone seriously interested in comparing a small number of films would actually look at the shape of the entire SL distribution, as well as less quantifiable aspects of style, before, possibly, selecting specific summary statistics to highlight aspects of stylistic similarities or differences they wish to draw attention to. I infer from this that the main value in using summary statistics with quantified SL data is that it aids the comparison of reasonably large bodies of films, as a starting point for undertaking more detailed analysis rather than being an end in itself. Even here it seems necessary, if the data are available, to 'screen out' films whose SL distributions do not admit simple summary.

The above is a way of saying that I think that debates about the relative merits of the median and ASL may be misplaced; as a minimum you need both, or their equivalent (but see later). If you must have the debate, arguments from robustness against the ASL because of the effects of outliers are – if intended to be general – unsustainable at a practical level. I think a more careful distinction between clear outliers and what is to be expected in the tails of samples from skewed distributions is needed, but regardless of this the empirical analyses in the paper suggest that outliers/extreme values either often have little numerical effect on ASL calculations, or that the larger absolute changes have limited impact on the relative rankings of films and comparative judgements that would be made about them. That a small number of outliers, unless very seriously extreme, will have little effect on ASL calculations if n is at all large, and of the order typical since, say, the 1970s is mathematically obvious. It is, in any case, a simple matter to check if claimed outliers do have much of an effect on the ASL, without relying on generalised assertions about lack of robustness.

There are examples in the literature where what I would view as a substantial minority of observations (over 10%) in a long right tail of an SL distribution, with no obviously serious discontinuities, are treated as 'outliers'. Omitting such cases can affect the ASL noticeably but not the median. I think such analyses are based on a misconception of what constitutes an outlier. Even allowing for the practice, my perhaps naive thought is that a long tail of, say, 10% of an SL distribution is an important and even interesting aspect of its structure (is this the same as 'style'?), and the fact that the median is designed to ignore it is not obviously a commendation[6].

I'd suggest, despite the attention given to it in the literature, that the issue of outliers as a *statistical* problem is something of a 'red herring'. From a 'filmic' viewpoint – if this is a word – and as I think Barry Salt has suggested, uncontentious outliers may be among the more interesting aspects of a film's 'style'.

[6] I'm ducking the issue of how you define the tail of a distribution, but I hope the thought is clear. The closest to a definition I've seen in the cinemetric literature, and where the 10% figure comes from, is [16], where what I'd call a tail is defined as beginning at an arbitrary distance from the median of the data dictated, I think, by the software used. The paper uses the criterion to identify what are called 'outliers' which I think is a misuse of the term. The arbitrariness of the definition is not a problem; any definition of what constitutes the tail of a skewed distribution will be arbitrary to some extent.

Ignoring outliers, arguments against the use of the ASL because SL distributions are not 'bell-shaped' – that is, they are skew – are misconceived. You can't summarise a skew distribution, in terms of measures of location, using just a single statistic. The mode, median and mean (ASL) do different jobs and, if you must have just one it is 'fitness for purpose' that should determine the choice. I'll return to this shortly, but first want to indulge in an idea that recurs every time I read papers advocating the median in preference to the ASL.

It will be obvious that I am largely unimpressed by the statistical arguments that have been made for preferring the median to the ASL in the cinemetric literature. Where not based on unsupported and, as I would argue, unsustainable assertion, the empirical demonstrations which my statistical tastes more naturally incline me to can be interpreted quite comfortably in an opposite manner to that intended by their authors. Taking a more positive view though, it is intriguing to ask *why* there is this possible 'groundswell' of support for the median, even if the support can be ill-articulated. Is there a 'primal urge' or yearning, among some scholars at least, for a statistic that captures some *absolute* essence of 'style', whatever 'style' is. Where other than implicit (and I think subtly different from the 'median is to be preferred because it is the median' argument) 'style' or one aspect of it seems to be associated with terms like 'typical' or where the 'mass' of data lies.

That is, 'style' is being associated with some idea of 'typicality' and then the median is being proposed as a/the statistic that captures this aspect of style. Since the ASL is not 'typical' in the way that most people think of as typical it follows that the median is a more appropriate absolute measure of style than the ASL. I can't resist suggesting – not tongue-in-cheek – that following this line of thought, the mode or modal class seems a better choice than the median but it seem to have few champions. I won't pursue the reasons why this might be the case.

I'm not really sure if the above argument makes any sort of sense, and if it does I'm not entirely convinced that it does differ that much from the 'median is to be preferred because it is the median' argument. If it does, the 'subtle' difference is that the 'rightness' of the median and not the ASL as a quantifiable measure of style proceeds from purely 'filmic' considerations. Statistics has nothing to say about the issue.

I'm acting as 'devil's advocate' here. My view (one I think has been repeatedly stressed by Barry Salt, though I don't have the references to hand to confirm this) is that the main purpose of quantification is to aid (objective) *comparison*. I conceive of this as a 'relativistic' endeavour, so 'absolutist' considerations are not that relevant. The important issue though, regardless of the coherence of this thought, and I think at the heart of one of Yuri Tsivian's concerns, is that if you are at all seduced by arguments in favour of the median, whatever they are, should you be worried that all the effort that has been invested in calculating ASLs has been wasted – as implied in Nick Redfern's comments included in Yuri's statement of the problem?

My answer would be 'No', there is no serious cause for worry. Taking caveats about distributional shape as read, and notwithstanding the repeated insistence above that you can't generally reduce the description of a skewed distribution, in terms of measures of location, to a single statistic, the median and ASL will tend to work equally well for comparative puposes.

This is suggested empirically by the generally good linear relationship between the median and ASL demonstrated for 1920s silent films, early sound films, and Hollywood films from 1935 and particularly those from around 1970. The rank ordering of films is very similar whatever statistic is used and, even allowing for the effect of extremes, differences that exist do not look great enough to affect any broad comparisons one might wish to make.

This can be explained in various ways. Very informally, the smaller the SLs tend to be in general the more 'squashed up' the distribution will be towards the left, and the smaller both the median and ASL will be. A strong relationship between the two might then be expected. This argument would not work if outliers did have a severe effect on the ASL – the main plank in many arguments against the ASL – but the results in this paper suggest that severe effects are the exception rather than the rule.

Mathematically, if a functional relationship existed between μ and σ or, most simply, between μ_L and Ω_L of the form $\Omega_L = \beta\mu_L$, then summaries, in the ideal case of a lognormal distribution, using just a single statistic are possible. In some statistical applications there are theoretical reasons for expecting such a relationship, but none that I'm aware of for SL distributions.

More realistically, the relevant plots in the paper suggest that an empirical and *statistical* model of the form $\Omega_L = \beta\mu_L + \varepsilon$, where ε is an error term, is plausible. The effect of ε will be to induce 'churn' into the rank order of films obtained from the two statistics but nothing too serious if the variance of ε is not large. Outliers, such as *Applause* for the early sound films, will be manifest on bivariate plots of the median against ASL. The natural statistical thing to do is omit such outliers is estimating the parameter, β, that characterises the general relationship between the median and ASL, and investigate separately why a case is an outlier (rather than simply saying it is better described by the median or ASL).

Another useful statistical concept is that of a *surrogate variable*. The idea occurs in disciplines as various as econometrics, chemometrics, psychometrics, etc. (sometimes under different names) where a surrogate variable 'stands in' for some other variable that may be difficult or expensive to acquire, or may be unobservable. The requirement is that the surrogate variable be sufficiently strongly correlated with the variable one would ideally like to use, that analyses based on it will lead to similar conclusions.

The idea might be applied to film data in various ways. One is that 'style' is an unobservable variable and that either the median or ASL are suitable surrogate variables for it. The other, of more direct relevance to Yuri's question, is that if you think the median is the ideal measure of style to use but only have ASL measurements, the latter is a perfectly adequate surrogate variable to substitute for the former.

There will be circumstances when the ASL is not a suitable descriptor of style, but if you accept the argument that this will not usually be attributable to outliers despite what is claimed in the literature, this is likely to be for distributions where the median is also unsuitable.

References

[1] Adams, B., Dorai, C. and Venkatesh, S. (2002) *Formulating film tempo: the computational media aesthetics methodology in practice*, in C Dorai and S Venkatesh (eds.) *Media Computing: Computational Media Aesthetics*. Norwell, MA: Kluwer Academic Publishers: 57-84.

[2] Barnett, V. and Lewis, T. (1994) *Outliers in Statistical Data, 3rd ed.*, Wlley, Chichester.

[3] Cutting, J.E., Delong, J.E. and Nothelfer, C.E. (2010) *Attention and the evolution of Hollywood film*, Psychological Science 21, pp. 440-447.

[4] Cutting, J.E., Brunick, K.L. and Delong, J.E. (2011a) *The changing poetics of the dissolve in Hollywood film*, Empirical Studies of the Arts, pp. 149-169.

[5] Cutting, J.E., Brunick, K.L. and Delong, J.E. (2011b) *How act structure sculpts shot lengths and shot transitions in Hollywood film*, Projections 5 (2011), pp. 1-16.

[6] Cutting, J.E., Brunick, K.L. and Delong, J.E. (2012) *On shot lengths and film acts: A revised view*, Projections 6, pp, 142-145.

[7] Cutting, J.E., Brunick, K.L., Delong, J.E., Catalina, C. and Candan, A. (2011) *Quicker, faster, darker: Changes in Hollywood film over 75 years*, i-Perception 2, pp. 569-576.

[8] 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, pp. 115-125.

[9] Delong, J.E., Brunick, K.L. and Cutting, J.E. (2012) *Film through the human visual system: finding patterns and limits*, in *The Social Science of Cinema*, J.C. Kaufman and D.K. Simonton, eds., Oxford University Press, New York, in press.

[10] Grubbs, F.E. (1969) *Procedures for detecting outlying observations in samples*, Technometrics 11, pp. 121.

[11] Hawkins, D.M. (1980) *Identidication of outliers*, Chapman and Hall, London.

[12] Kang, H-B. (2003) *Affective contents retrieval from video with relevance feedback*, in TMT Sembok, HB Zaman, H Chen, S Urs, SH and Myaeng (eds.) *Digital Libraries: Technology and Management of Indigenous Knowledge for Global Access*. Berlin: Springer: 243-252.

[13] Redfern, N. (2009a) *The impact of sound technology on the distribution of shot lengths in motion pictures*, http://nickredfern.files.wordpress.com/2009/09/nick-redfern-the-impact-of-sound-technology-on-hollywood-film-style1.pdf

[14] Redfern, N. (2009b) *Measures of central tendency in Cinemetrics*, http://nickredfern.wordpress.com/2009/12/10/measures-of-central-tendency-in-cinemetrics/

[15] Redfern, N. (2012a) *The lognormal distribution is not an appropriate parametric model for shot length distributions of Hollywood films*, http://nickredfern.files.wordpress.com/2012/02/nick-redfern-the-lognormal-distribution-and-hollywood-cinema.pdf

[16] Redfern, N. (2012b) *Robust time series analysis of ITV news bulletins*, http://nickredfern.files.wordpress.com/2012/04/nick-redfern-robust-time-series-analysis-of-itv-news-bulletins.pdf

[17] P. Rousseeuw and A. Leroy (1987) *Robust Regression and Outlier Detection*, Wiley, New York.

[18] Schaefer, R, and Martinez, T.J. (2009) *Trends in network news editing strategies from 1969 through 2005*, Journal of Broadcasting and Electronic Media 53 (3): 347-364.

[19] Vasconcelos, N. and Lippman, A. (2000) *Statistical models of video structure for content analysis and characterization*, IEEE Transactions on Image Processing, 9 (1): 3-19.

[20] Venables, W.N and Ripley, B.D. (2002) *Modern Applied Statistics with R: Fourth Edition*, Springer, New York.

Appendix - the lognormal distribution

The shape of the distribution of a variable X is lognormal – call it $L(\mu, \sigma)$ – if it can be described mathematically as

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

If X is lognormal its logarithm, $Y = \ln X$, has the normal distribution, $N(\mu, \sigma)$, described mathematically as

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

Both distributions are completely described given knowledge of just two parameters (μ, σ).

Commonly, and for observed data and not just idealised distributions, *summary statistics* are used to identify characteristics of a distribution of interest, often for comparative purposes, a distinction usually being made between measures of *location* and measures of *scale*, with *shape* also being a concern in some contexts.

The most familiar measures of location are the mean, median and mode; measures of scale include the standard deviation (or its square, the variance) and the inter-quartile range. The notation $(\mu_N, \Omega_N, \Phi_N)$ can be used for the mean, median and mode of the Normal distribution. Since

$$\mu_N = \Omega_N = \Phi_N = \mu$$

notational refinement can be dispensed with and μ used for all three quantities.

Things are not so simple for the lognormal distribution; using L rather than N as a subscript to distinguish the two distributions

$$\mu_L = \exp(\mu + \sigma^2/2) \quad \Omega_L = \exp(\mu) \quad \Psi_L = \exp(\mu - \sigma^2)$$

and

$$\Psi_L < \Omega_L < \mu_L.$$

Here, $\exp(\bullet)$ is the exponential function, and we can also write $\mu = \ln \Omega_L$.

In principle any pair of statistics that involves μ and σ can be used for summary purposes; (Ω_L, μ_L) (the median and ASL) is one possibility. Neither is separately useful unless there is a fairly strong relationship between them, the one because it involves only one of (μ, σ), the other because it conflates them.

For skewed distributions not so simply described mathematically it may be possible to get by with two summary statistics; more may be needed; or none (of those commonly used) may be suitable. Whay you can't do is get by with one.