Distributional Thinking about Film Style: Quantile comparisons of motion picture shot length data

Nick Redfern

Independent scholar, United Kingdom
Correspondence: nickredfernres@outlook.com
ORCID: 0000-0002-7821-2404

Peer review: This article has been subject to a double-blind peer review process

Abstract

This article illustrates the use of quantiles as a means of describing and comparing motion picture shot length distributions. This approach is conceptually and computationally simple and leads us to think distributionally about shot lengths rather than focusing on individual values. The result is a better understanding of how this element of film style of two (or more) films differs.

Keywords: Computational film analysis; statistical literacy; film editing; shot length distribution; quantiles

Introduction

In this article, I demonstrate the use of quantiles as a conceptually and computationally simple approach to describing and comparing shot lengths in motion pictures. This approach has been overlooked to date in the quantitative analysis of film style, but it is one that results in a better understanding of how the editing style of two (or more) films differs.

This article is aimed at researchers who wish to apply quantitative methods to shot length data derived from motion pictures and introduces the core concepts of distributional thinking about shot length data using quantiles. I include technical details so that readers can understand the derivation of the statistics presented, but a quantile-based approach is a simple and intuitive way of describing and comparing shot lengths. It is no more conceptually difficult than the average shot length (ASL) but avoids the methodological pitfalls of the latter. I present two case studies that illustrate using quantiles to compare shot lengths in two films and in two groups of films. An online tutorial is available for those who would like to apply the quantile-based methods demonstrated here (see Supplementary material).

Computational Analysis of Motion Picture Editing

Computational film analysis (CFA) is a field of inquiry within the digital humanities that aims to understand the formal properties of the cinema (**Burghardt, et al., 2020; Heftberger, 2018**), and exists as a network of interconnected systems weaving together a specialist domain knowledge of a phenomenon of interest in the humanities (the cinema), knowledge about the design, execution, and validation of research projects, competencies in the use of quantitative methods to collect and analyse data, and competencies in the application of computational tools to process that data (**Redfern, 2020a**). Computational film analysis is more than the statistical analysis of film style. Although CFA employs statistical methods, it inherits a broader outlook on what is possible with quantitative methods and computational tools from data science to embrace exploratory data analysis, statistical modelling, machine learning, data visualization, and computer programming to tell the story of the data. CFA falls within the scope of greater data science described by David Donoho (**2017**), which comprises the tasks of data gathering, preparation, and exploration; data representation and transformation; computing with data; data modelling; data visualization and presentation; and science about data science. Applications of CFA cover a wide range formal and stylistic elements of motion pictures, including scripts (**Del Vecchio, et al., 2021**), dialogue (**Hołobut & Rybicki, 2020**), social networks (**Weng et al. 2009**), colour (**Chen, et al., 2012**), shot scales (**Svanera, et al., 2019**), visual content (**May & Shamir 2019**), documentary 'voice' (**Villanueva Baselga, et al., 2021**), and sound design (**Redfern, 2020b**).

A common application of quantitative methods to the analysis of film style is to determine if the duration of shots in two (or more) motion pictures or two (or more) groups of films differ, and, if so, by how much. This has been applied in a wide variety of cases. Cutting, et al **(2010)** analysed the shot lengths of 150 Hollywood films and showed an increasing tendency of shot lengths to cluster into sequences of shorter and longer takes over time from 1935 to 2005. Redfern (**2020c**) identified differences in the durations of shots in classical and post-classical Hollywood and showed that animation films are stylistically distinct from other genres irrespective of when they were produced. Schaefer and Martinez (**2015**) tracked changes in shot duration in US television news from 1969 to 2005; and Redfern (**2014a**) compared shot lengths in British television news in relation to broadcast time and content. Kim and Lee (**2020**) analysed shot lengths and their relationship to emotion in Korean television series; while Butler (**2014**) analysed shot duration in American television sitcoms, reporting a statistically significant difference between shot lengths in single-camera and multi-camera productions. Baxter, et al. (**2017**) compared shot lengths in the films of Mack Sennett, Charlie Chaplin, and D. W. Griffith, identifying differences in style between filmmakers whose use of editing evolved over time.

One approach to comparing shot durations in motion pictures is to ask:

> *Question 1: how does the typical shot length in film X compare to the typical shot length in film Y?*

Answers to this question are typically presented as a comparison of the films' respective cutting rates as measured by their average shot lengths (ASLs), which describes the mean waiting time between cuts (**Salt, 1974, 1992**). The size of the difference between average shot lengths of *X* and *Y* is conventionally interpreted as the difference in style between *X* and *Y*. This is the dominant approach used in statistical analyses of film style and film scholars such as Salt (**1992**), Bordwell (**2002**), Buckland (**2006**), O'Brien (**2009**), Roggen (**2019**), and Vyas and Shekhawat (**2021**), amongst many others. All rely on the ASL as a means of describing differences in editing style – and in many cases, exclusively so. The Cinemetrics database (http://www.cinemetrics.lv) led by Yuri Tsivian (**2009**) aims to reveal patterns in editing over time and between different groups of films by comparing the average shot durations in a database containing shot length data on over twenty thousand motion pictures (including films, television programmes, adverts, etc.). A key problem with this approach is that differences in ASLs do not necessarily reflect differences in style. For example, Barry Salt challenges Andrew Sarris's claim that Lewis

Milestone's *The Front Page* (1931) is edited more quickly than *His Girl Friday* (1940), arguing that this is not in fact the case because both films have the same ASL whilst also pointing out they are not stylistically similar: 'The average shot length of both movies is the same; however, the Milestone film achieves this by having a larger number of very short shots and a larger number of very long shots' (**Salt, 1974: 18**).

An alternative way to address this problem is to systematically compare all the shot lengths in two films and to ask:

> *Question 2: do the shots in film X tend to be longer than shots in film Y?*

Redfern (**2014b**) described a dominance statistics approach to answering this question, using Cliff's *d* to describe the extent to which shots in one film are likely to be of longer duration than shots in another films, and the Hodges-Lehmann Difference (HLD), which is the median of the pairwise differences between the shots in two films, to estimate the size of this difference in seconds. These statistics describe global differences between the duration of shot lengths in films and do not identify the nature of these differences, though this can be addressed by use of the empirical cumulative distribution function as a graphical method for comparing all the shot lengths in the two films to identify where differences lie.

In both of the above cases, it was necessary to refer to the distribution of shot lengths to contextualise the meaning of the summary measures used. It therefore makes sense to begin any comparison of motion picture shot length by thinking *distributionally* about differences in film style and focussing on shot length data as a collective entity rather than individual data values. We can therefore ask:

> *Question 3: how do shot lengths in specific parts of their respective distributions compare between films?*

The rest of this article demonstrates a quantile-based approach to answering this question.

Distributional Thinking

A *data set* comprises a collection of pieces of related information produced by measuring some properties of a group of objects. A data set is characterised by *variation*, which, in simple terms, is the quality of a measured property of an object to vary (**Makar & Confrey, 2005**). Chris Wild argues that the need for statistics flows from variation: 'the statistical response to variation is to investigate, disentangle, and model patterns of variation in order to learn from them. Virtually all of the ways statisticians do this involve looking at the data through a lens which is distribution' (**Wild, 2006: 21**). A *distribution* is a representation of the variation of a data set that allows us to organise and examine data efficiently to gain an overall understanding of how the data varies. Aisling Leavy argues that an understanding of distribution requires 'an awareness of the propensity of a variable to vary and comprehension of how that variability contributes to the notion of the distribution as an aggregate rather than a collection of individual data points' (**Leavy, 2006: 90**). *Distributional thinking* is quantitative reasoning about variation, distribution, and the relationship between them (**Bakker & Gravemeijer, 2004; Prodromou, 2007**).

When we talk about the distribution of a data set, we need to describe a range of features, including the overall shape of the distribution and any deviations from the overall pattern. To describe a distribution, we need to ask the following questions about its various features:

- *centre*: where is the mass of the data located? Where is the centre of the data located? What is the typical value of a data set?
- *spread*: how much variability is there in the data set?
- *symmetry (skewness)*: is the distribution symmetrical? Is the bulk of the data to the left of the distribution with a long right tail (positively skewed)? Or is the bulk of the data to the right of the distribution with a long left tail (negatively skewed)?
- *modality*: how many peaks does the distribution have?
- *peakedness*: is the shape of the peak(s) flat and broad or tall and pointed?
- *tailedness (kurtosis)*: how much of the data is located in the tails of the distribution relative to the centre?
- *outliers*: are there any deviations from the overall pattern of the data? Are there observations that are noticeably distinct from the bulk of the data?

In attending to these features, we attempt to account for the variation in a data set that deals with its complexity that may arise in a range of different situations.

In talking about the editing of motion pictures, a data set comprises the duration of each shot in the film – the variable of interest – and a *shot length distribution* is the way in which we think and talk about that data and how we compare different data sets for different films. However, if we look at the common applications of statistics to questions of film editing, we see that concepts of 'variation' or 'distribution' are seldom present. Distributional thinking of the sort Wild and Leavy describe as fundamental to statistical reasoning about data is rarely a part of the statistical analysis of film style. At present, most descriptions in the literature on film style do not address the features of shot length distributions, relying on comparisons of ASLs alone. Most researchers applying statistical methods to the analysis of style in the cinema in fact collect no data and never produce a shot length distribution, relying instead on dividing the running time of a film by the number of shots. Consequently, they are unable to provide any information about the shot length data for a film beyond the ASL. Even when researchers do collect the full data set on shot durations for a film, they only report the ASL and ignore other features of shot length distributions that are potentially interesting, such as the variability of the data or the shape of the shot length distribution. Only a small proportion of the literature addresses features of shot length distributions beyond the ASL (see, for example, **Baxter et al., 2017; Fujita, 1989; Kohara and Niimi, 2013**; and **Redfern, 2020c**). The result is that a lack of distributional thinking characterises the most applications of quantitative methods to the analysis of film style.

Quantiles

A conceptually simple method of describing and comparing shot length distributions is to use the quantiles of the distributions. A *quantile* (Q_p) is a cut point dividing a data set arranged in order from the smallest value to largest so that a specified proportion p of the data set lies below that point (see **Altman and Bland, 1994**). The p-th quantile of a data set is found using the quantile function

$$Q_p = \{x: Pr(X \leq x) = p\}$$

The quantile function is the inverse of the empirical cumulative distribution function, which is the probability that the duration of a shot is less than or equal to some specified value. These functions are different representations of the same information, but, for the purposes of analysing shot length distributions, the quantile function is preferable because its output is expressed in terms of the information that interests the researcher – the p-th quantile of a data set is x seconds – rather than as a probability. Commonly used quantiles are the median ($Q_{0.5}$) of a data set, dividing the range into two equal parts, or the lower ($Q_{0.25}$) and upper quartiles ($Q_{0.75}$) that cut off the lower and upper 25% of a data set, respectively, but quantiles for any value of p can be used. There is one less quantile than the numbers of groups created by dividing a data set into subsets of equal size: to divide a data set into 20 equal parts we need 19 quantiles. In calculating a set of quantiles for a data set a 'quantile profile' is produced that summarises that data set (**Johnson et al., 2015**), and which can then be used to compare two or more data sets. The use of quantiles to systematically compare two distributions can be implemented via a shift function that plots the difference between the differences between the quantile profiles of the distributions (see **Doksum, 1974; and Rousselet, et al., 2017**).

Descriptive statistics are 'indices' of a distribution (Leavy 2006), summarizing data sets using a small number of features that make large data sets manageable. The features of a shot length distribution listed above can be described in terms of its quantiles. The median shot length is a measure of location, while the interquartile range (the difference between the upper and lower quartiles: $Q_{0.75} - Q_{0.25}$) describes the spread of a distribution. The symmetry of a distribution is described by the skewness coefficient,

$$S = \frac{Q_{0.25} + Q_{0.75} - 2Q_{0.5}}{Q_{0.75} - Q_{0.25}}$$

which takes on values between -1 and 1. Values of S greater than 0 indicate positive skewness, which is typical for motion picture shot length distributions. The kurtosis of a distribution can also be described in terms of quantiles:

$$T = \frac{(Q_{0.875} - Q_{0.625}) + (Q_{0.375} - Q_{0.125})}{Q_{0.75} - Q_{0.25}}$$

Kurtosis measures the combined weight of the mass of data in the tails of a distribution relative to its centre, with higher values of kurtosis indicating there are a lot of data points in the tails. For *T*, the two terms in the numerator measure the combined weight of the shoulders of a distribution while the denominator is the IQR, which describes the middle of the distribution. The terms in the numerator will be large if relatively more data is located in the shoulders than in the centre of a distribution resulting in higher values of *T* (**Moors, 1988**). Taken together, these four statistics – media, interquartile range, quantile skewness, and quantile kurtosis – provide an informative, intuitive, and robust numerical summary of the distribution of shot lengths in a motion picture. These numerical descriptions do not require any assumptions to be made about possible models for a shot length distribution.

There are multiple methods for calculating the quantiles of a data set. Here I use the Harrell-Davies estimators produced using the hdquantile function in the Hmisc package (**Harrell, 2021**) for the statistical programming language R (**R Core Team, 2021**).

Comparing shot lengths in two films using quantile differences

To illustrate the use of quantiles when comparing the shot length distributions of two films, I analyse two Laurel and Hardy films: *You're Darn*

Tootin' (1928) and *Hog Wild* (1930). **Table 1** summarises the distributions numerically and **Figure 1** presents the kernel density plots of the shot lengths distributions in these films. From **Figure 1** we see that the distribution for *Hog Wild* is more positively skewed, which is confirmed by its larger quantile skewness, and has a sharper, higher peak compared to the broader, flatter peak of than that of *You're Darn Tootin'*. It also has more mass concentrated in the tails of its distribution: this can be seen in **Figure 1** where the density of the tails is greater than in *You're Darn Tootin'* and is also indicated by its greater quantile kurtosis in **Table 1**. Although these films have similar interquartile ranges, *Hog Wild* has a greater range due to the presence of shorter shots in the lower tail and longer shots in the upper tail. A key difference between these distributions is therefore the difference between their respective tails. Overall, we see that shot lengths in *Hog Wild* are more varied than those in *You're Darn Tootin'*; and that *Hog Wild* tends to be edited more quickly, with more shots of shorter duration and a high density of shots at approximately 2 seconds duration compared to *You're Darn Tootin'*, whilst at the same having shots of greater duration. *You're Darn* Tootin' is less diverse stylistically, with shots concentrated more evenly within a narrower range of lengths. Interestingly, **Figure 1** shows both films have a bump in the upper tail at around 24 seconds that is not captured by any of the numerical summaries, and which may be of interest. This illustrates the importance of using graphical displays when talking about shot length distributions.

If we ask question 1 – how does the typical shot length in *You're Darn Tootin'* compare to the typical shot length in *Hog Wild*? – we find that these two films both have an ASL of 6.6 seconds, which is interpreted as there being no difference in cutting rate according to the conventional use of ASLs. However, **Figure 1** shows that while these films may have the same ASL, they have different shot length distributions indicating there are differences in the style of these films that are not captured by any of the statistics commonly used to compare shot lengths.

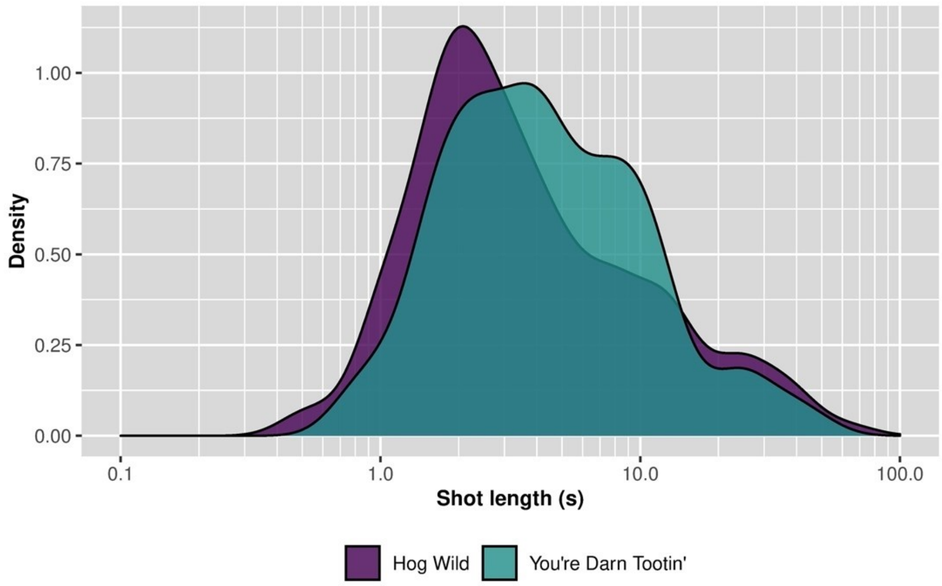
Turning to question 2 – do the shots in *You're Darn Tootin'* tend to be longer than shots in *Hog Wild*? – we find that by calculating the pairwise differences between every shot in these films by subtracting the length of each shot in *You're Darn Tootin'* from the length of each of shot in *Hog Wild*, the Hodges Lehman difference between the distributions is HLD = -0.6s and Cliff's d = -0.13, indicating the duration of shots in *Hog Wild* tends to be slightly shorter overall than those in *You're Darn Tootin'*. These statistics have captured an aspect the difference between these shot length distributions at a global level; however, they provide no information about other differences in which we might be interested. Like the ASL, they tell us nothing about the shape of the distributions and do not tell us how specific parts of the distributions of these two films differ.

To answer question 3 – how do shot lengths in specific parts of their respective distributions compare between *You're Darn Tootin'* and *Hog Wild*? – I plot the quantile profiles for each film (**Figure 2.A**) and the difference between the quantiles of each film (**Figure 2.B**), subtracting the quantiles of *You're Darn* Tootin' from those of *Hog Wild* so that negative differences indicate quantiles for which *You're Darn Tootin'* are greater and positive differences identify quantiles of *Hog Wild* are greater. To simplify this example, I have limited the number of quantiles to 19, ranging from $Q_{0.05}$ to $Q_{0.95}$ and increasing by increments of 0.05, but we could choose any number of quantiles. When we look at the differences of the quantiles of the shot length distributions of these two films, the nature of the differences between the two distributions is immediately apparent. For most quantiles, shot duration in *You're Darn Tootin'* is greater than in *Hog Wild*: this is clear in **Figure 2.A** where the quantile profile of *You're Darn Tootin'* is higher than that of *Hog Wild*. Above quantile $Q_{0.8}$ shots in *Hog Wild* tend to be longer than those in *You're Darn Tootin'*, and in Figure 2.A we see that the quantile profiles have crossed over so that the profile for *Hog Wild* is now above that of the other film.

Table 1: Statistical summary of two Laurel and Hardy films

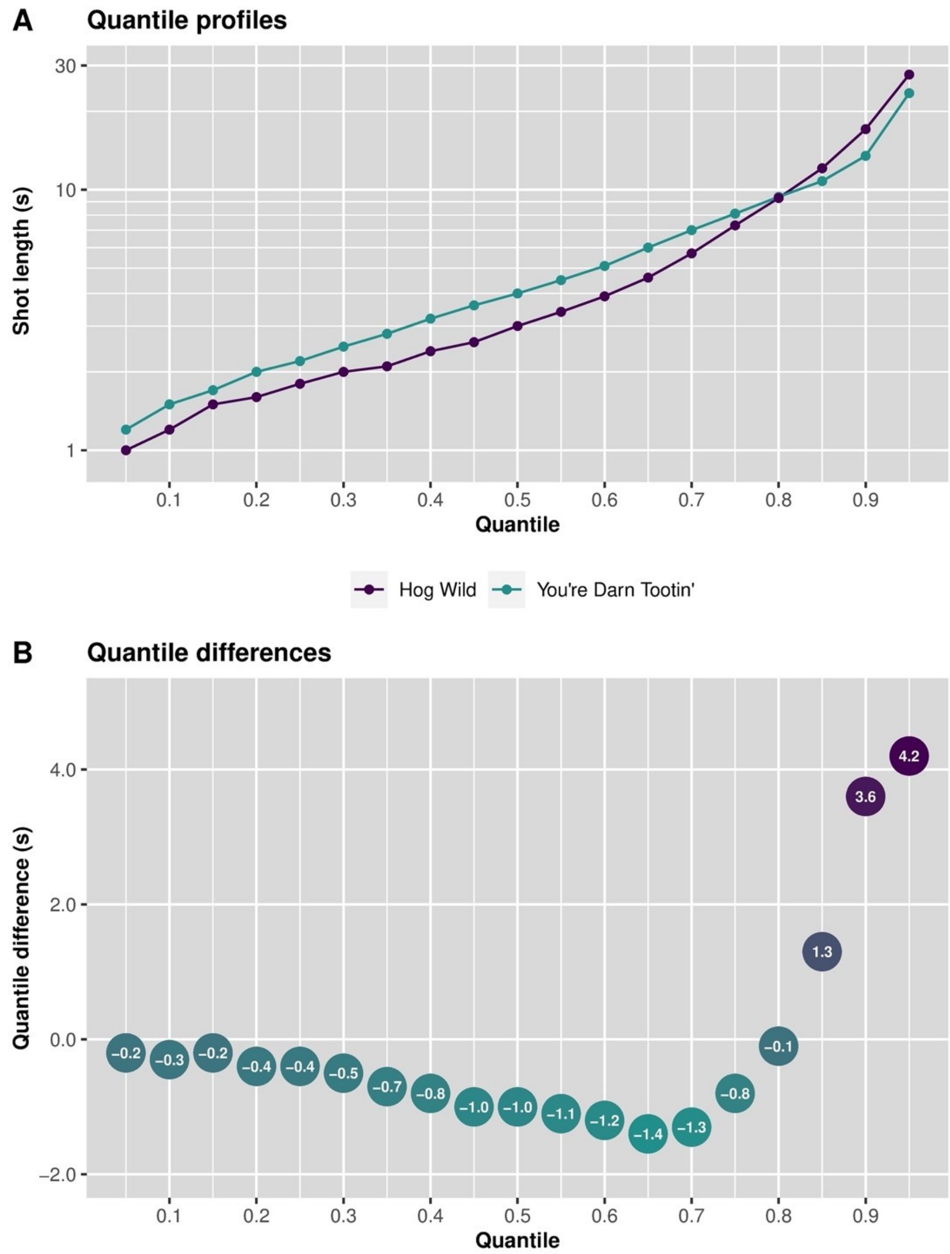
	You're Darn Tootin'	*Hog Wild*
Shots (N)	189	169
Mean (s)	6.6	6.6
Minimum (s)	0.7	0.5
Lower quartiles (s)	2.2	1.8
Median (s)	4.0	3.0
Upper quartile (s)	8.0	7.2
Maximum (s)	49.1	65.1
IQR (s)	5.8	5.4
Quantile skewness	0.39	0.57
Quantile kurtosis	1.28	1.93

Figure 1: Kernel density estimates of shot length distributions of two Laurel and Hardy films: You're Darn Tootin' (1928) and Hog Wild (1930).



If we want to ask the question, 'How do shot lengths in *You're Darn Tootin* compare to those of *Hog Wild*?', then comparing the quantiles of the shot length distributions of these films shows that the difference in editing in these films is more complicated than can be conveyed by comparing their average shot lengths. In this example, it is not clear what the fact these two films have the same ASL means given the differences in their style indicated by their shot length distributions. The dominance statistics approach provides a more accurate global description of the differences in shot lengths of these films but does not automatically lead us to consider the nature of those differences. A key advantage of the quantile approach is that by using this simple method we can identify and talk about the complicated nature of these differences by thinking about shot lengths *distributionally*.

Figure 2: Quantile comparison of shot length distributions in two Laurel and Hardy films: You're Darn Tootin' (1928) and Hog Wild (1930). (A). The quantile profiles for each film. (B).



Note: Quantile differences - negative differences in the above image indicate quantiles for which shots in *You're Darn Tootin'* tend to be of greater duration and positive differences identify quantiles when shots in *Hog Wild* tend to be longer.

Comparing Shot Length Distributions in Two Groups of Films

The quantile method described above can be easily extended to comparing groups of films. To illustrate the comparison of shot lengths distributions of two groups of films based on quantiles I analyse the shot length data for four silent (*The Ring* (1927), *The Farmer's Wife* (1928), *Champagne* (1928), and *The Manxman* (1929)) and five sound films

(*Blackmail* (1929), *Murder!* (1930), *The Skin Game* (1931)*, Rich and Strange* (1931), *Number Seventeen* (1932)) directed by Alfred Hitchcock between 1927 and 1932. I collected shot length data from the *Early Hitchcock* PAL DVD release of these films and corrected the duration of each shot to 24 frames per second by multiplying by a factor of 1.041667. I removed the opening and closing credits from each film, but all other titles are included.

From the quantile-based descriptive statistics in **Table 2**, we see that the four silent films (*The Ring*, *The Farmer's Wife*, *Champagne*, and *The Manxman*) have shot length distributions that are relatively consistent, with similar median shot lengths and spread. With the shift to sound filmmaking in 1929, Hitchcock's early sound films (*Blackmail*, *Murder!*, *The Skin Game*) show an obvious change in editing style with increases in the median as shots tended to become longer in duration and more varied as seen in the change in the interquartile range. There is also a change in the shape of shot length distribution as indicated by the increase in quantile skewness and quantile kurtosis. The greater part of the difference between these early sound films and the silent movies that preceded them is an increase in the spread of shot lengths above the median as dialogue shots required longer takes, while the spread of shots below the median shot length remains largely unchanged. For the later sound films we see a shift to an editing style characterised by shorter takes similar to his silent films but with shot length distributions that are more skewed and kurtose like his first sound films. *Rich and Strange* has a number of rapidly edited montage sequences, such as the Paris sequence or the leisure activities aboard the cruise ship, and a series of drawn-out conversational sequences that are cut more slowly, while *Number Seventeen* is largely comprised of a rapidly cut extended chase sequence in between two slower cut sequences at the house and the harbour, and which again maintains a similar distinction between dialogue-heavy scenes and action.

The summary statistics give us an overall impression of the difference in shot length distributions between Hitchcock's late silent and early sound films, but we do not yet know anything about the nature of that difference. Visualising the distributions by plotting their kernel densities in **Figure 3** gives concrete meaning to the median shot length and interquartile range, as indicated by the quantile lines of each plot, making it clear how the descriptive statistics relate to data and illustrating how shot lengths in *Blackmail*, *Murder!*, and *The Skin Game* are more widely dispersed than those of the other films and how shot lengths initially increase before shortening in duration. The shape of the distribution of shot lengths for *Blackmail* in **Figures 3** and **4.A** resembles the distributions of the silent film in the quartiles of the lower tail and the quartiles of the upper tail of the sound films that followed it. This pattern emerges due to the unique production circumstances of *Blackmail*, which was released in both silent and sound versions described by Charles Barr as 'works of continuously inventive *bricolage*. Juxtaposing them scene by scene, one registers a set of permutations: points at which variously, (a) both versions use 'silent' visuals; (b) both versions use 'sound' visuals; (c) silent and sound visuals are mixed within a scene; (d) the two films use entirely different visuals' (**Barr, 1983: 123**). The change in distributional shape indicated by the quantile skewness and kurtosis values is evident, with the similarity of shape of the silent films clear to see and the shift in the mass of the data for the sound films in the upper tail of the distribution relative to the centre of the distribution.

While the kernel densities in **Figure 3** make it clear how the shape of the distributions in these films have changed, the quantile profiles in **Figure 4.A** make it easier to see where the distributions differ and the size of those differences. The quantile profiles clearly show that the silent films are much more consistent in the distribution of their shot lengths than the sound films, which exhibit much greater variation in shot length at different quantiles. We can also see evidence of a hybrid editing style: with the exception of one of the sound films (*Blackmail*), Hitchcock's editing style after the introduction of sound meant that takes of shorter duration were slightly shorter than those of his silent films and, at the same time, longer takes in Hitchcock films increased in duration. To compare the quantiles across two groups of films, I subtract the value of the *p*-th quantile of each silent film from the *p*-th quantile of each sound film. With a sample containing four silent films and five sound films we have a total of twenty differences for each quantile, and **Figure 4.B** plots the resulting difference distributions in which a negative difference indicates that the *p*-th quantile for a silent film is greater than the *p*-th quantile for a sound film and a positive difference indicates a greater shot length at the *p*-th quantile in a sound film. From **Figure 4.B** we see that differences at the lower quantiles ($Q_{0.05}$ to $Q_{0.35}$) are centred around negative values, reinforcing the fact that the shorter takes in Hitchcock's sound films tend to be shorter than those in his silent films. At the same time, we see that the distributions of the differences between the upper quartiles lie to the right of zero seconds again showing that the longer takes in the sound films tend be of greater duration than the silent films. As the distribution of differences shows two peaks for quantiles in the range $Q_{0.55}$ to $Q_{0.85}$ we can also identify the presence of sub-groups within the sample, which is accounted for by the second change in Hitchcock's editing style with *Rich and Strange* and *Number Seventeen* and the shift of mass in the distributions to the lower tail that occurs with the use of rapidly edited montage and chase sequences.

Overall, we can see from the changes in the shape of the shot length distributions of his late silent and early sound films that Hitchcock's editing style changed in two ways with the introduction of synchronised sound. The distribution of shot lengths became much more polarised in Hitchcock's first sound films as shorter shots became shorter and longer takes became longer. However, in the later sound films *Rich and Strange* and *Number Seventeen*, we see a shift to a more rapid editing style with the use of shorter takes in non-dialogue sequences that were not only shorter than those of the early sound films but also shorter than those we see in Hitchcock's late silent films.

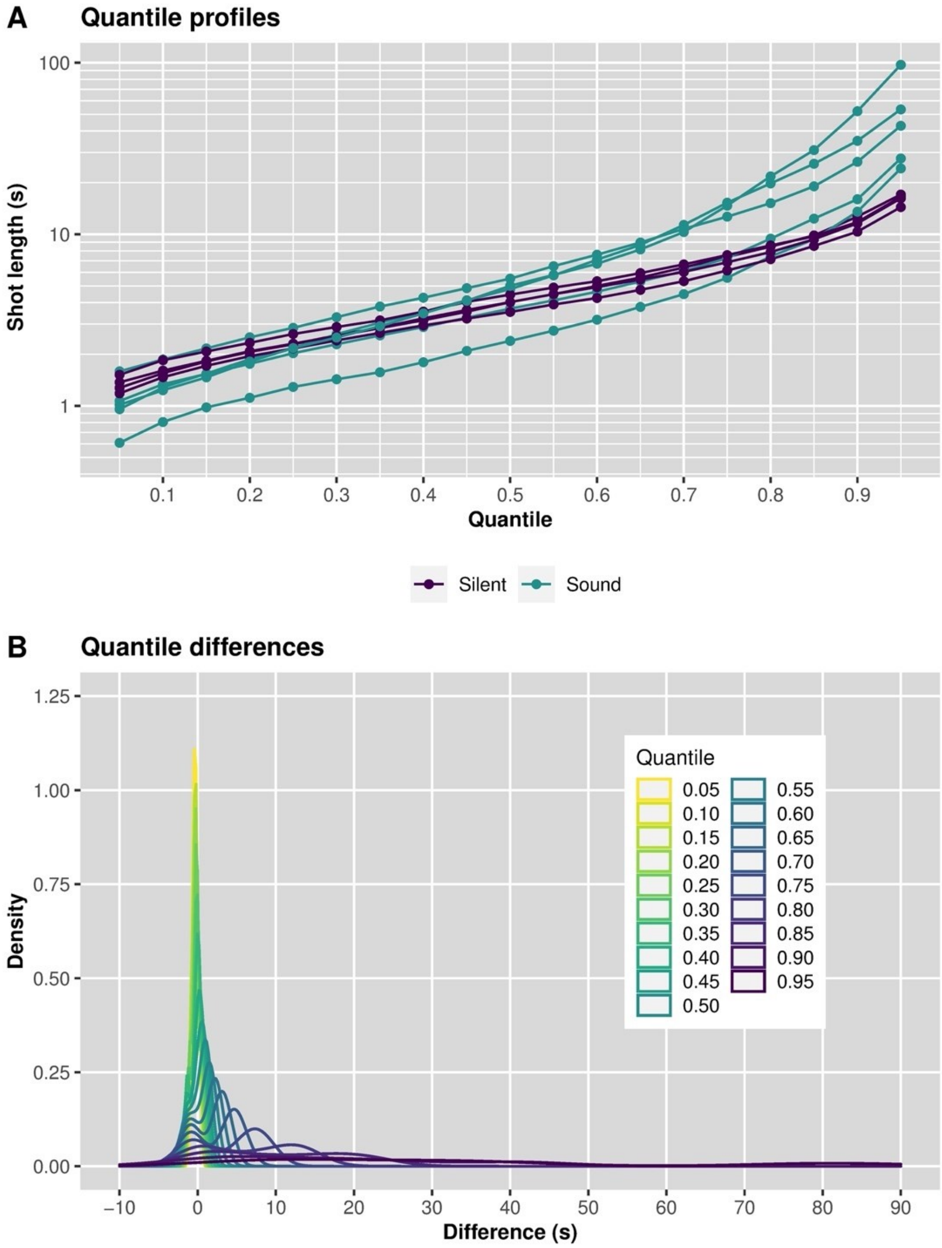
Table 2: Quantile statistics of shot length distributions of films directed by Alfred Hitchcock, 1927-1932

	The Ring (1927)	The Farmers Wife (1928)	Champagne (1928)	The Manxman (1929)	Blackmail (1929)	Murder (1930)	The Skin Game (1931)	Rich and Strange (1931)	Number Seventeen (1932)
Shots (N)	1056	1007	893	808	438	438	269	687	655
Length (s)	5286.62	5786.39	5253.81	4908.10	5025.42	6013.50	4856.66	4885.40	3706.00
Mean (s)	5.01	5.75	5.88	6.07	11.47	13.73	18.05	7.11	5.66
Minimum (s)	0.04	0.25	0.04	0.60	0.92	0.50	0.80	0.20	0.10
Lower quartile (s)	2.16	2.30	2.28	2.62	2.85	2.17	2.17	2.03	1.29
Median (s)	3.53	4.02	4.02	4.45	5.51	4.79	5.01	3.68	2.39
Upper quartile (s)	6.15	6.87	7.38	7.56	12.67	15.28	14.69	7.26	5.57
Maximum (s)	57.63	78.83	63.71	52.50	148.17	223.90	281.00	79.10	77.70
IQR (s)	3.99	4.57	5.10	4.94	9.82	13.11	12.52	5.24	4.28
Quantile skewness	0.32	0.25	0.32	0.26	0.46	0.60	0.55	0.37	0.49

Quantile kurtosis	1.48	1.39	1.40	1.30	1.63	1.86	2.60	1.93	1.94

Figure 3: Shot length distributions of films directed by Alfred Hitchcock, 1927-1932. The quantile lines in each density plot show the lower quartile (Q0.25), median (Q0.5), and upper quartile (Q0.75) of each distribution.

Figure 4: A) Quantile profiles of shot length distributions of films directed by Alfred Hitchcock, 1927-1932. (B) Difference distributions for pairwise differences between quantiles of shot length distributions of films directed by Alfred Hitchcock, 1927-1932.



Note: Positive differences in the images above indicate that quartiles the sound films in the sample are higher than those of the silent films.

Conclusion

In this article I demonstrated that the problem of comparing the duration of shots in two or more motion pictures can be approached from a range of different perspectives, but that whichever approach is adopted, distributional thinking is essential. However, at present distributional thinking about motion picture shot lengths remains uncommon due to overreliance on the ASL. Inferences about ASLs tell us nothing the distribution of shot lengths in motion pictures or about what differences in style may exist between films; while the common practice of counting the number of shots in a film and dividing by the running time of a film to calculate the ASL is, in effect, to put one's faith in statistics in the absence of data. Even those resources that do make shot length data available to researchers, such as the Cinemetrics database, report only a handful of summary statistics and rely heavily on the ASL.

A strictly quantitative approach will never be enough to analyse the style of a motion picture – not every question we wish to ask about style in the cinema requires the application of quantitative methods and not every element of film style is quantifiable. Statistical analyses of shot length distributions will only answer questions about shot length distributions and so the methods demonstrated here will typically be employed in a context that employs both qualitative and quantitative methods. But if film scholars are to apply statistical methods to questions of film style, those methods should illuminate our understanding and be methodologically sound. If our goal is to understand differences in shot lengths as differences in style between films, the quantile approach described in this article is simple and intuitive to understand and accurately describes the nature of the differences between these shot length distributions and the size of those differences. Most importantly, it is founded in distributional thinking.

Supplementary Material

A tutorial demonstrating how the summaries and plots presented in this paper were produced using the R statistical programming language is available for those who would like to use these methods in their own research. Access the tutorial here: https://rpubs.com/nr62_rp33/SL-quantiles.

Nick Redfern has been teaching and researching about film since 2001. He has taught film, media, and television studies at Manchester Metropolitan University, the University of Central Lancashire, and Leeds Trinity University. His work on computational film analysis has been published in the *Journal of Japanese and Korean Cinema*, *Statistica*, *Journal of Data Science*, *Post Script*, *Digital Scholarship in the Humanities*, *Umanistica Digitale*, *Acta Universitatis Sapientiae: Film and Media Studies*, *Research Data Journal for the Humanities and the Social Sciences*, and *Music, Sound, and the Moving Image*. You can read more about computational film analysis at https://computationalfilmanalysis.wordpress.com.

Tables

Images

References

Altman D. G., and Bland, J. M., 1994. Statistics notes: quartiles, quintiles, centiles, and other quantiles. *BMJ*, 309, p. 996. DOI: 10.1136/bmj.309.6960.996. [Accessed: 1 May 2021].

Bakker, A., and Gravemeijer, K. P., 2004. Learning to reason about distribution. *In* Ben-Zvi, D. and Garfield, J., eds., *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*. Springer: Dordrecht, pp. 147-168.

Barr, C., 1983. *Blackmail*: silent and sound. *Sight and Sound*, 52 (2), pp. 123-126.

Baxter, M., Khitrova, D., and Tsivian, Y., 2017. Exploring cutting structure in film, with applications to the films of DW Griffith, Mack Sennett, and Charlie Chaplin. *Digital Scholarship in the Humanities*, 32 (1), pp. 1-16. DOI: 10.1093/llc/fqv035. [Accessed: 1 May 2021].

Bordwell, D., 2002. Intensified continuity visual style in contemporary American film. *Film Quarterly*, 55 (3), pp. 16-28. DOI: 10.1525/fq.2002.55.3.16. [Accessed: 25 February 2022].

Buckland, W., 2006. *Directed by Steven Spielberg: Poetics of the Contemporary Hollywood Blockbuster*. London: Continuum.

Burghardt, M., Heftberger, A., Pause, J., Walkowski, N.-O., and Zeppelzauer, M., 2020. Film and video analysis in the digital humanities – an interdisciplinary dialog. *Digital Humanities Quarterly*, 14 (4). Available at: http://www.digitalhumanities.org/dhq/vol/14/4/000532/000532.html. [Accessed: 25 February 2022].

Butler, J., 2014. Statistical analysis of television style: what can numbers tell us about TV editing? *Cinema Journal*, 54 (1), pp. 25–44. DOI: 10.1353/cj.2014.0066. [Accessed: 25 February 2022].

Chen, I.-P., Wu, F.-Y., and Lin, C.-H., 2012. Characteristic color use in different film genres. *Empirical Studies of the Arts*, 30 (1), pp. 39–57. DOI: 10.2190/EM.30.1.e. [Accessed: 25 February 2022].

Cutting, J. E., DeLong, J. E., and Nothelfer, C. E., 2010. Attention and the evolution of Hollywood film. *Psychological Science*, 21 (3), pp. 432–439. DOI: 10.1177/0956797610361679. [Accessed: 25 February 2022].

Del Vecchio, M., Kharlamov, A., Parry, G., and Pogrebna, G., 2021. Improving productivity in Hollywood with data science: Using emotional arcs of movies to drive product and service innovation in entertainment industries. *Journal of the Operational Research Society* 72 (5), pp. 1110-1137. DOI: 10.1080/01605682.2019.1705194. [Accessed: 25 February 2022].

Doksum, K., 1974. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Annals of Statistics*, 2 (2), pp. 267-277. DOI: 10.1214/aos/1176342662. [Accessed: 1 May 2021].

Donoho, D., 2017. 50 years of data science, *Journal of Computational and Graphical Statistics*, 26 (4), pp. 745-766. DOI: 10.1080/10618600.2017.1384734. [Accessed: 25 February 2022].

Fujita, K., 1989. Shot length distributions in educational TV programs. *Bulletin of the National Institute of Multimedia Education*, 2, pp. 102-116.

Harrell Jr, F. E., with contributions from C. Dupont and many others., 2021, Hmisc: Harrell Miscellaneous. R package version: 4.5-0. https://CRAN.R-project.org/package=Hmisc. [Accessed 1 May 2021].

Heftberger, A., 2018. *Digital Humanities and Film Studies: Visualising Dziga Vertov's Work*. Cham: Springer.

Hołobut, A. and Rybicki, J., 2020. The stylometry of film dialogue: pros and pitfalls. *DHQ: Digital Humanities Quarterly*, 14 (4). Available at: http://www.digitalhumanities.org/dhq/vol/14/4/000498/000498.html. [Accessed: 25 February 2022].

Johnson, W. D., Beyl, R. A., Burton, J. H., Johnson, C. M., Romer, J. E., and Zhang, L., 2015. Use of Pearson's Chi-Square for testing equality of percentile profiles across multiple populations. *Open Journal of Statistics*, 5 (5), pp. 412–420. DOI: 10.4236/ojs.2015.55043. [Accessed: 1 May 2021].

Kim, Y., and Lee, S., 2020. Exploration of the characteristics of emotion distribution in Korean tv series: common pattern and statistical complexity. *IEEE Access* 8, pp. 69438–69447. DOI: 10.1109/ACCESS.2020.2985673. [Accessed: 25 February 2022].

Kohara, I., and Niimi, R., 2013. The shot length styles of Miyazaki, Oshii, and Hosoda: a quantitative analysis. *Animation*, 8 (2), pp. 163-184. DOI: 10.1177/1746847713487472. [Accessed: 1 May 2021].

Leavy, A. M., 2006. Using data comparison to support a focus on distribution: Examining preservice teacher’s understandings of distribution when engaged in statistical inquiry. *Statistics Education Research Journal*, 5 (2), pp. 89-114.

Makar, K., and Confrey, J., 2005. Variation talk: articulating meaning in statistics. *Statistics Education Research Journal*, 4 (1), pp. 27-54.

May, C., and Shamir, L., 2019. A data science approach to movies and film director analysis, *First Monday*, 25 (6). DOI: 10.5210/fm.v24i6.9629. [Accessed: 25 February 2022].

Moors, J. J. A., 1988. A quantile alternative for kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 37 (1), pp. 25-32. DOI: 10.2307/2348376. [Accessed: 1 May 2021].

O’Brien, C. 2009. Sous les toits de Paris and transnational film style: an analysis of film editing statistics. *Studies in French Cinema*, 9 (2), pp. 111-125.

Prodromou, T., 2007. The coordination of distributional thinking. Paper presented at the 56th session of the International Statistical Institute. Lisbon, Portugal. Available at: https://iaseweb.org/documents/papers/isi56/CPM80_Prodromou.pdf [Accessed 1 May 2021].

R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. [Accessed 1 May 2021].

Redfern, N. 2014a. The structure of ITV news bulletins. *International Journal of Communication* 8, pp. 1557–1578. Available at: https://ijoc.org/index.php/ijoc/article/viewFile/1646/1151. [Accessed 1 May 2021].

Redfern, N., 2014b. Comparing the shot length distributions of motion pictures using dominance statistics. *Empirical Studies of the Arts*, 32 (2), pp. 257-273. DOI: 10.2190/EM.32.2.g. [Accessed 1 May 2021].

Redfern, N. 2020a. What is computational film analysis? Available at: https://computationalfilmanalysis.wordpress.com. [Accessed 25 February 2022].

Redfern, N., 2020b. Sound in horror film trailers. *Music, Sound, and the Moving Image*, 14 (1), pp. 48-71. DOI: 10.3828/msmi.2020.4. [Accessed 25 February 2022].

Redfern, N., 2020c. Modelling shot lengths of Hollywood cinema with the Dagum distribution. *Statistica*, 80 (1), pp. 81-99. DOI: 10.6092/ISSN.1973-2201/9910. [Accessed 25 February 2022].

Roggen, S., 2019. Cutting rates and shot scales in 1950s CinemaScope: a systematic style analysis, *Screening the Past*, 44. Available at: http://www.screeningthepast.com/2019/03/cutting-rates-and-shot-scales-in-1950s-cinemascope-a-systematic-style-analysis/. [Accessed 25 February 2022].

Rousselet, G. A., Pernet, C. R., and Wilcox, R. R., 2017. Beyond differences in means: robust graphical methods to compare two groups in neuroscience. *European Journal of Neuroscience*, 46 (2), pp. 1738-1748. DOI: 10.1111/ejn.13610. [Accessed: 1 May 2021].

Salt, B., 1974. Statistical style analysis of motion pictures. *Film Quarterly*, 28 (1), pp. 13-22. DOI: 10.2307/1211438. [Accessed: 1 May 2021].

Salt, B., 1992. *Film Style and Technology: History and Analysis*, 2nd edition. London: Starword.

Schaefer, R., and Martinez, T., 2009. Trends in network news editing strategies from 1969 through 2005. *Journal of Broadcasting & Electronic Media*, 53 (3), pp. 347–364. DOI: 10.1080/08838150903102600. [Accessed 25 February 2022].

Svanera, M., Savardi, M., Signoroni, A., Kovacs, A. B., and Benini, S., 2019. Who is the film's director? Authorship recognition based on shot features. *IEEE MultiMedia* 26 (4), pp. 43-54. DOI: 10.1109/MMUL.2019.2940004. [Accessed 25 February 2022].

Tsivian, Y., 2009. Cinemetrics, part of the humanities' cyberinfrastructure. *In*: Ross, M., Manfred Grauer, M., Freisleben, B., eds., Digital *Tools in Media Studies: Analysis and Research – An Overview*. Bielefeld: Transcript, pp. 93-100. DOI: 10.25969/mediarep/2558. [Accessed 25 February 2022].

Villanueva Baselga, S., Perez-Montoro, M., and Sánchez, L., 2021. The use of statistics to define taxonomies of film voices: A practical approach to documentary classification. *Visual Communi*cation, 20 (2), pp. 248–270. DOI: 10.1177/1470357219850591. [Accessed 25 February 2022].

Vyas, N., and Shekhawat, S., 2021. Statistical style analysis of Hindi biopics: exploring the genre conventions in recent years. *Quarterly Review of Film and Video*, pp. 1-20. DOI: 10.1080/10509208.2021.1980364. [Accessed 25 February 2022].

Weng, C.-Y., Chu, W.-T., and Wu, J.-L., 2009. RoleNet: movie analysis from the perspective of social networks. *IEEE Transactions on Multimedia*, 11 (2), pp. 256-271. DOI: 10.1109/TMM.2008.2009684. [Accessed 25 February 2022].

Wild, C., 2006. The concept of distribution. *Statistics Education Research Journal* 5 (2): 10-26.

To cite this article:

Redfern, N., 2022. Distributional thinking about film style: quantile comparisons of motion picture shot length data. *Exchanges: The Interdisciplinary Research Journal*, 10(1), 21-42. Available at: https://doi.org/10.31273/eirj.v10i1.853.