Robust measures of scale for shot length distributions

Nick Redfern

1. Introduction

The distribution of shot lengths in a motion picture can be analysed statistically (Salt 1974), but to date the methods employed have been vulnerable to fundamental errors arising from the two characteristic features of this data – its positive skew and the presence of outliers. Commonly cited statistics of film style such as the average (mean) shot length and the standard deviation do not accurately describe the style of a film due to these factors – the mean is pulled away from the mass of the data under the influence of outliers and does not locate the centre of the distribution, while the standard deviation is based upon the mean is itself subject to the influence of outliers. Removing outliers from shot length data is undesirable, because although shots of extended duration (such as the opening shot of *Touch of Evil* [1958])) are relatively rare in the cinema they nonetheless represent the aesthetic decisions of filmmakers. In order to retain all the data for a film without arriving invalid conclusions, robust statistical methods that are unaffected by the presence of extreme values or deviations from assumptions about nature of the data should be employed (Rousseeuw 1991, Daszykowski *et al*. 2007). For example, the median shot length provides a simple robust alternative to the mean because it will locate the centre of any distribution irrespective of its shape and is not affected by the presence of outliers in the data. In order to understand the style of a film we need to understand the variation of the data, and so the median shot length needs to be accompanied with a robust measure of scale in place of the standard deviation. This paper looks at some of the robust alternatives to the standard deviation that may be applied in film studies.

2. Robust measures of scale

The robustness of an estimator (θ) may be expressed by its breakdown point and its influence function. The breakdown point is the proportion of data that can be given arbitrary values before θ becomes arbitrarily bad. The influence function measures the impact of a data point on the estimator, and is unbounded if θ goes to infinity for an arbitrarily large x and bounded if it remains within the range of the original data. It is desirable that an estimator has a high breakdown point and bounded influence function.

The standard deviation (σ) is not a robust estimator of scale: its breakdown point is 0% and its influence function is unbounded, meaning that just a single outlier can make σ an arbitrarily bad statistic of film style. In fact, σ is especially vulnerable to the presence of outliers as it is calculated using the squared distance of a data point from the mean and this gives added weight to data points in the tails of the distribution at the expense of the mass of the data. An additional problem is that σ is defined by its relationship to a measure of location (i.e. the mean), and is unsuitable for asymmetric distributions. Six alternatives to the standard deviation are considered as robust measures of scale for the statistical analysis of film style.

Mean absolute deviation

The mean absolute deviation (d_n) is the average absolute distance of every shot length from a measure of location. Any measure of location may be used, but the median shot length is used here:

$$d_n = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - med_j(X_j)\right|$$

d_n is a superior measure of scale to the standard deviation when the assumptions required for σ are not met (Gorard 2004); but it is also adversely affected by outliers with a breakdown point of 0% and an unbounded influence function.

Median absolute deviation

The median absolute deviation (MAD) is the median of the absolute deviations of each shot length from the median shot length:

$$MAD = med_i\left(\left|X_i - med_j(X_j)\right|\right)$$

The median absolute deviation has the highest breakdown point possible (50%) and its influence function is bounded (but not smooth), making it a robust estimator in the presence of outliers. The key drawback of MAD is that it is based on the median as a measure of location, where $med_i \pm MAD$ includes approximately 50% of the data, and is therefore most suitable when dealing symmetrical distributions.

However, the median absolute deviation can still be used for analysing film style. For example, using the coefficient of median deviation (the ratio of MAD to the median shot length) it is possible to sort the late-silent and early-sound films of Alfred Hitchcock into three groups by k-means clustering: the purely silent films (M = 0.47), the two versions of *Blackmail* (M = 0.60) that combine silent and sound visuals (Barr 1983), and the pure sound films (M = 0.72). These groupings are presented in Figure 1.

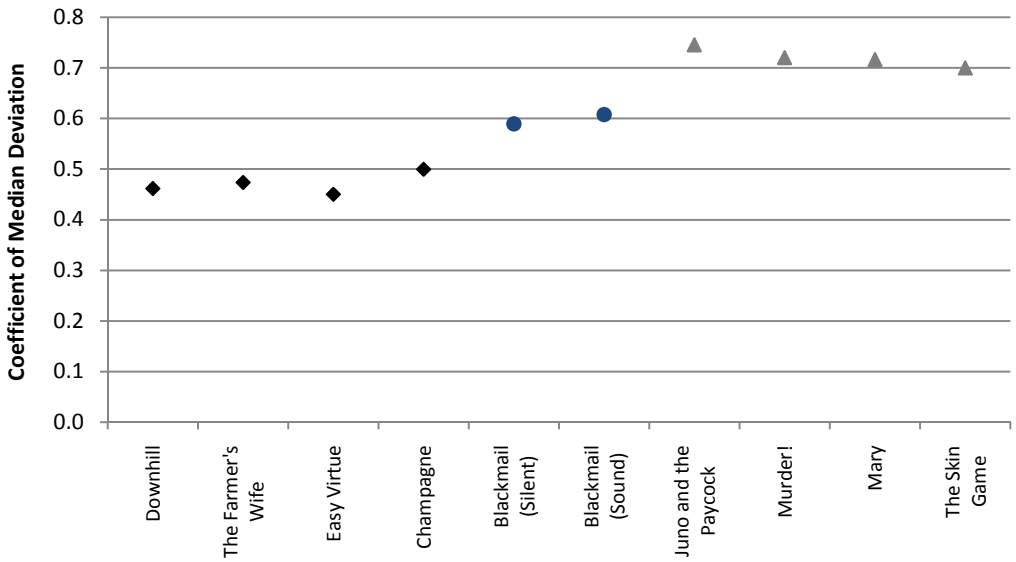

Figure 1 Coefficient of median deviation for silent and sound films directed by Alfred Hitchcock, 1927 to 1931. Source: Cinemetrics database.

Gini's mean difference

Gini's mean difference is the expected value of $\binom{n}{2}$ absolute pairwise differences between shot lengths:

$$G = \frac{1}{\binom{n}{2}} \sum \left| X_i - X_j \right| \; ; i < j \; .$$

The major advantage of G over the mean and median absolute deviations is that it does not depend on some measure of location, and is therefore appropriate for asymmetric distributions. However, G is not resistant to outliers, with a breakdown point of 0% and an unbounded influence function. Consequently, G is prey to the same weaknesses as the standard deviation.

S_n* and *Q_n

Croux and Rousseeuw (1992) and Rousseeuw and Croux (1993) proposed two measures of scale as alternatives to the median absolute deviation and Gini's mean difference. The first estimator, S_n, is the median of the n medians of the absolute differences between shot lengths:

$$S_n = c_{Sn} \times 1.1926 \; med_i \left\{ med_j \left| X_i - X_j \right| \; ; \; j = 1, 2, \dots, n \right\} ,$$

where the factor 1.1926 is for consistency, and c_{Sn} is bias correction factor based on the sample size equal to $n/(n - 0.9)$ if n is odd or to 1 if n is even. The second estimator, Q_n, is

$$Q_n = c_{Qn} \; \times 2.2219 \left\{ \left| X_i - X_j \right| \; ; \; i < j \right\}_{(k)} ,$$

where the factor 2.2219 is for consistency, and $k = \binom{h}{2}$ and $h = [n/2] + 1$. Q_n is the kth order statistic of the $\binom{n}{2}$ absolute pairwise differences between shot lengths, and which for large n is equal to the lower quartile of these values. c_{Qn} is bias correction factor based on the sample size, and is $n/(n + 1.4)$ if n is odd or $n/(n + 3.8)$ if n is even.

Both estimators have a breakdown point of 50% and a bounded influence function (which for Q_n is smooth). As neither estimator is dependent upon a measure of location and calculates the distance of each shot length from every other shot length, they are appropriate for asymmetric distributions typically encountered in the cinema. The major drawback is that these estimators may only be available as part of expensive statistical software packages many film scholars will not have access to; and to calculate either S_n or Q_n step-by-step (using Excel, for example) is a simple but labour-intensive process.

Interquartile range

The interquartile range (IQR) is defined as the distance between the lower (Q_1) and the upper (Q_3) quartiles of the data:

$$IQR = Q_3 - Q_1$$

The IQR is also less robust than other measures mentioned here, with a breakdown point of 25%, but it is resistant to outliers due to its focus on the centre of the distribution. At the same time, we learn nothing about the tails of a distribution as the IQR only provides information about the middle 50% of the data. The interquartile range is simple to calculate and easy to understand, but may be misleading when dealing with asymmetric distributions

as the distance between the quartiles and the median may not be equal. (For this reason, the semi-interquartile range should not be used for shot length distributions). These problems may be overcome by ensuring that the distribution of shot lengths in a film is described by the five number summary of the minimum shot length, the lower quartile, the median shot length, the upper quartile, and the maximum shot length; and by using box-plots to represent the data.

3. The dispersion of shot lengths in three short films of Laurel and Hardy

To compare the robust measures of scale discussed above, frame-accurate shot length data was collected from three Laurel and Hardy short films produced for the Hal Roach Studio: *Chickens Come Home* (1931), *Scram!* (1932), and *Busy Bodies* (1933). The summary statistics for each film are presented in Table 1.

The distribution of shot lengths in each of the films included here is represented in the box-plots in Figure 2. Box-plots are an excellent method for conveying a large amount of information about a data set quickly and simply; and are an effective method of comparing multiple data sets. From analysing the box-plots for shot lengths motion pictures we can compare the location and variation of the data, and identify the skew and the presence of outliers. The core of the data is defined by the box, which covers the distance between the lower and upper quartiles (i.e. the IQR), and the horizontal line within the box represents the median value of the data. Extending from the box, the error bars make the inner fences beyond a data point is considered an outlier. An outlier is defined here as greater than $Q_3 + (IQR \times 1.5)$ and the error bars extend to largest value within this limit; while an extreme outlier as greater than $Q_3 + (IQR \times 3)$. Typically, there are no outliers at the low end of a shot length distribution (although this may not be the case when transforming the data), and the error bar descends to the value of the shortest shot.

From the information in Table 1 and Figure 2, it is clear that all three films share the characteristics that are typical of the shot length distributions of motion pictures: that is, they are positively skewed and have a number of outlying data points.

Table 1 Summary statistics of three Laurel and Hardy short films

	Chickens Come Home	***Scram!***	***Busy Bodies***
Length (s)	1771.4	1193.7	1112.4
Shots	227	198	174
Mean shot length (s)	7.8	6.0	6.4
Standard deviation (s)	7.8	6.5	8.4
Skew	2.4	2.8	2.6
Minimum shot length (s)	0.5	0.8	0.5
Lower quartile (s)	2.4	2.3	1.7
Median shot length (s)	5.0	3.4	3.1
Upper quartile (s)	10.2	7.1	7.0
Maximum shot length (s)	52.0	42.7	47.6

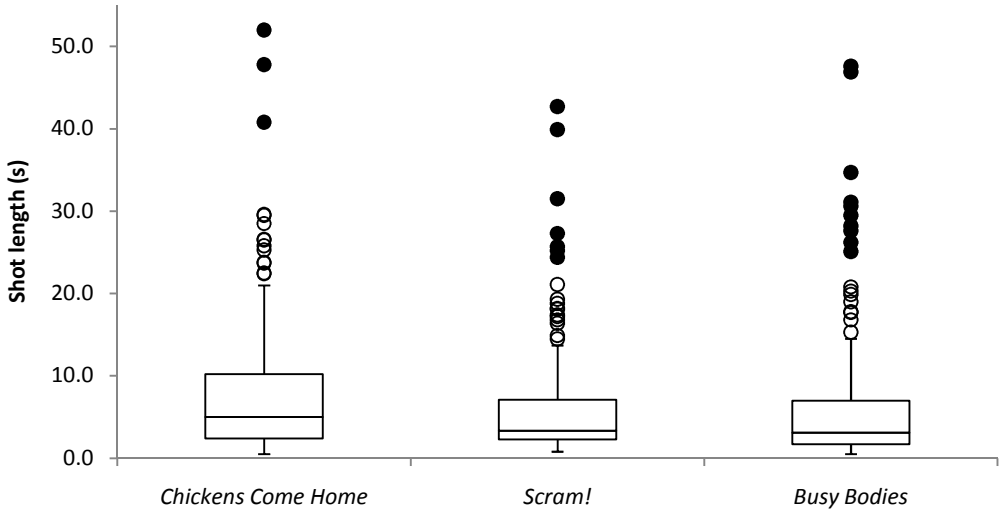

Figure 2 Shot length distributions in three Laurel and Hardy short films.

Using non-robust statistics of film style such as the mean shot length or the standard deviation will lead us to make erroneous inferences about the style of these films. For example, the standard deviations suggest that shot lengths in *Busy Bodies* are more widely dispersed than in *Scram!* However, we can see from the five number summary and the box-plots for these films that the dispersion of their shot lengths is very similar, and this argument is generally supported by looking at the robust measures of scale in Table 2. The median absolute deviation, S_n, and Q_n all indicate that the spread of the data for these two films are similar, as suggested by Figure 2. The mean absolute deviation also indicates that these films are more similar than σ would lead us to believe, but d_n is greater than these three estimators as it is subject to greater influence of outlying shot lengths. The interquartile range also leads us to conclude that these films are similarly dispersed, but it is only from looking at the five-number summaries and/or the box-plots that we can see how the distance between Q_3 and the median is much greater than between Q_1 and the median. An exception is Gini's mean difference, which for these films indicates a difference in scale similar to that for the standard deviations; and, when interpreted in the context of the information given by the five-number summary and the box-plots, clearly gives a mistaken impression that shot lengths in *Busy Bodies* are more dispersed than in *Scram!*

Table 2 Robust measures of scale for three Laurel and Hardy short films

	Chickens Come Home	***Scram!***	***Busy Bodies***
Mean absolute deviation (s)	5.2	3.7	4.6
Median absolute deviation (s)	3.2	1.4	1.7
Gini's mean difference (s)	7.5	5.7	7.3
S_n (s)	4.4	2.2	2.5
Q_n (s)	4.2	2.2	2.4
Interquartile range (s)	7.8	4.8	5.3

To take a second example, the standard deviation of *Chickens Come Home* is greater than that of *Scram!* but less than that of *Busy Bodies*; and we might interpret this as showing that the style of this film differs from the others in different ways. However, this is clearly a flawed inference: from the five-number summary and the box-plots we can see that shot lengths in *Chickens Come Home* are more dispersed than in *Scram!* and *Busy Bodies*. From the robust statistics in Table 2, we can see that d_n, the median absolute deviation, S_n, Q_n, and the interquartile range all lead us to the correct interpretation about the differences in style of these films. As before, Gini's mean difference gives a misleading interpretation of the data, indicating that *Chickens Come Home* and *Busy Bodies* have similarly dispersed shot lengths and that these two films differ from *Scram!* in a similar manner when this is clearly not the case.

4. Conclusion

The asymmetric nature of shot length distributions and the presence of outliers in shot length data require the application of robust statistical methods. Appropriate robust measures of scale for the statistical analysis of film style are the estimators S_n and Q_n, and the interquartile range (when used in the context of the five-number summary and in conjunction with box-plots). The median absolute deviation can be used to analyse film style, but it must be kept in mind that this estimator is based on a measure of location when analysing asymmetric shot length distributions. The mean absolute deviation can be a useful measure of film style, and its simplicity and intuitive nature are attractive; but there are other robust statistics that serve the same purpose with higher breakdown points and bounded influence function rendering d_n largely redundant. Gini's mean difference should not be used at all. It should also be clear that the use of graphical techniques such as box-plots play an important role in the exploratory data analysis of shot length distributions, and can prevent the analyst from arriving at mistaken conclusions about a film's style.

References

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

Croux C and Rousseeuw PJ 1992 Time-efficient algorithms for two highly robust estimators of scale, *Computational Statistics* 1: 411-428.

Daszykowski M, Kaczmarek K, Vander Heyden Y, and Walczak B 2007 Robust statistics in data analysis – a review: basic concepts, *Chemometrics and Intelligent Laboratory Systems* 85: 203-219.

Gorard S 2004 Revisiting a 90-year-old debate: the advantages of the mean deviation, British Educational Research Association Annual Conference, University of Manchester, 16-18 September 2004: http://www.leeds.ac.uk/educol/documents/00003759.htm, accessed 15 July 2010.

Rousseeuw PJ 1991 Tutorial to robust statistics, *Journal of Chemometrics* 5: 1-20.

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

Salt B 1974 Statistical style analysis of motion pictures, *Film Quarterly* 28 (1): 13-22.