Chapter 5

Graphical analysis – basics

In this chapter the sequence of SLs in a film is ignored in the analysis. A variety of different kinds of plot have been used in the cinemetric literature, of which the histogram is the most common. Commands needed for the basic plots are illustrated first, followed by more detailed discussion. *A Night at the Opera* is used for illustration until further notice, copied to `z` as in the previous chapter.

5.1 Histograms

5.1.1 Basics – an example

Histograms are perhaps the most widely used method for graphical presentation in cinemetrics. There are issues with their construction and interpretation, an initial illustration of which is provided in Figure 5.1. The histograms were obtained from the commands[1].

```
hist(z)
hist(z, 50, xlab = "SL", main = " ")
hist(z[z < 120], 50, xlab = "SL (< 120)", main = " ")
hist(z[z < 37], 100, xlab = "SL (< 37)", main = " ")
```

The top-left histogram uses the defaults in `R`. It is not that useful; the main problem is that, to accommodate the extreme value, rather large bin-widths of length 25 are imposed by the default.

The top-right histogram specifies the number of bins (or cells) to aim for, 50 in this case, and has the effect of reducing the bin-width to 5. The result is better but the scale is still affected by the extreme value[2]. The `xlab` and `main` arguments show how to control the labelling of the x-axis and title.

An obvious expedient to remove the effect of the outlying value is to omit it. In the bottom-left histogram this was done by selecting all SLs less than 120 using `z[z < 120]`. A bin-width of 2 results.

Something similar is done in the final plot, where a subset of the histogram is magnified by selecting only SLs less than 37 and increasing the number of bins so that the bin-width is 0.5. This choice was made in order to try and emulate Figure 3 in DeLong *et al.* (2012)[3].

The general appearance of the histogram for *A Night at the Opera* is characteristic of SL distributions. It is skewed with a single main peak and a long-tail. This kind of regularity,

[1] The layout shown was obtained within `R` by sandwiching the code between `par(mfrow = c(2,2))` and `par(mfrow = c(1,1))`. The first command switches to a 2 × 2 plotting grid, and the second switches back to the defaults. It is usually most convenient to place the commands in a function (Section 4.3).

[2] You don't necessarily end up with exactly the number of bins specified, since `R` 'works' at providing 'pretty' (and sensible) boundaries to avoid the large number of decimal places a slavish adherence to a specified number of bins might impose.

[3] Why this range was chosen is not discussed. The figure shown here doesn't exactly reproduce theirs, the exact appearance depending on both interval boundaries and the plotting algorithm used.

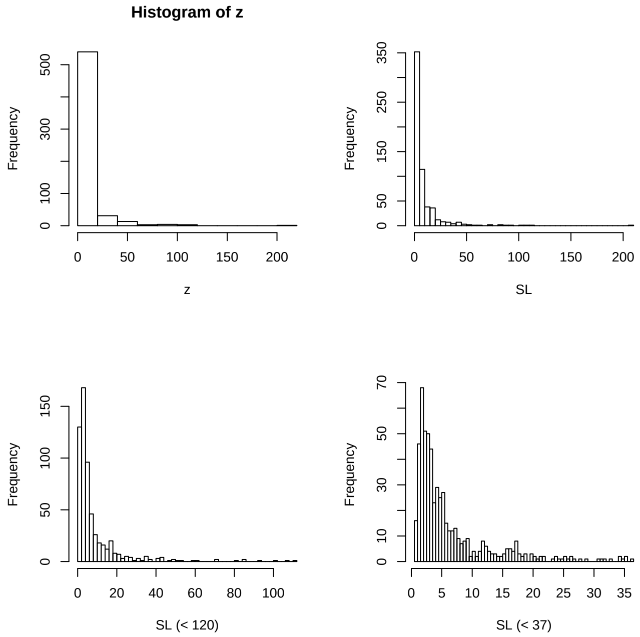

Figure 5.1: *Histograms for the SLs of* A Night at the Opera. *See the text for a discussion.*

noted by Salt (1974), led to the suggestion that many SL distributions might readily be modelled using a standard probability distribution. Later research led to the suggestion that the lognormal distribution was a fairly simple and appropriate model for, perhaps, 50% or more of films with average shot lengths (ASLs) of less than 15 seconds (Salt 2006, 2011).

This contention seems, variously, to have been both uncritically accepted, and strongly disputed, most notably by Redfern (2012a). For example, DeLong *et al.* (2012) present *A Night at the Opera* as an example of a film with a log-normal distribution whereas, as Redfern (2012a) demonstrates, it is not.

I don't intend to comment too much on this debate at this point. Given the generally similar qualitative appearance of many SL distribution the temptation to try and describe them with a common underlying model is understandable. The main point is that if the suitablity of the lognormal model is assessed on the basis of the appearance of a histogram, an element of subjective judgement is involved. This judgement is, in turn, influenced by the appearance of the histogram which is conditioned by the choices made in its construction. Initial 'choices' are usually made by the software. The user can usually modify them, with varying degrees of ease. It is actually much easier to do this in `R` than in some widely used speadsheet packages. The next section explores this.

5.1.2 Technicalities

A histogram is constructed as follows.

1. The range of the data is subdivided into a set of contiguous intervals ('cells' or 'bins' is other terminology used). The default, assumed here, is that the intervals are of equal width. The interval boundaries are determined by the *anchor point* (the position of the bottom-left corner of the histogram) which may extend slightly beyond the minimum data value.

   The choice of interval width (bin-width) *and* anchor point determines both the number of intervals and their boundaries. Equivalently, specifying the number of bins and an anchor point will determine the width[4].

2. The number of observations in each bin is counted. If observations fall exactly on a boundary a consistent rule should be applied to determine which side of the boundary they go.
3. The count, or frequency, in each bin is represented in a chart by a bar whose *area* is proportional to the frequency. If equal bin-widths are used the *height* of the bar is also proportional to frequency and this is what most users are familiar with. Adjacent bars should touch; charts purporting to be 'histograms' with visible gaps between bars are simply wrong[5].

In `R` Sturge's rule is the default for determining bin-widths in `hist(z)`. It is usually inadequate for SL data, where shorter intervals than it provides are desirable. Other rules are available but it is simplest, and legitimate, to experiment with the number of bins. Commands such as `hist(z, 50)` do this. More fully, this can be written as `hist(z, breaks = 50)`; the `breaks =` bit isn't needed if only a number is specified in the second position, but can be used if exact control over the interval boundaries is required.

It can be convenient to represent the histogram using a probability (density) rather than frequency scale. Using `hist(z, 50, freq = F)`, for example, does this (where `F` is shorthand for `FALSE`).

Try typing something like `NatO.hist <- hist(z, 50)`; the requested histogram appears. Type `NatO.hist` and a lot of information you mostly don't want appears. `NatO.hist` is an object that holds information about the histogram that can be extracted and manipulated if the urge to do so ovetakes you.

Type `names(NatO.hist)` to see what's available; you get

```
[1] "breaks"      "counts"      "intensities" "density"     "mids"
[6] "xname"       "equidist"
```

Typing `NatO.hist$breaks` brings up the break points (interval boundaries); `NatO.hist$counts` the counts in the bins; `NatO.hist$mids` the interval mid-points. These can usually be ignored but are there if needed.

5.1.3 Example continued - log-transformation

One problem I have with histograms like those in Figure 5.1 is a difficulty in making assessments about how lognormal the data look (if this is an issue). I disagree, for example, with Salt's (2006, 393) statement that for some of the examples he shows on p.392 the 'observed results fit well with a Lognormal distribution'. Similarly, as is about to be illustrated, the assertion of DeLong *et al.* (2012) that *A Night at the Opera* has a lognormal distribution is not sustainable.

If the SL data are lognormal (approximately) then they should be approximately normal after a logarithmic transformation (see the Appendix for the maths and the next section for `R` code). I suspect that most people find it easier to make assessments about normality than lognormality. Without worrying too much about the mathematics, a normal distribution should be symmetrical about a single peak – this is not a sufficient requirement for the data to be normally distributed, but forms a starting point for seeing if it isn't.

Figure 5.2 show histograms for logarithmically transformed SL data. The left-hand plot is the `R` default; the right-hand plot specified 30 as the number of bins. Part of the point here is to simply illustrate how the choice of numbers of bins (equivalently, bin-width) can affect the appearance of the histogram.

Neither plot looks convincingly normal, particularly that to the left. That is, the assertion that *A Night at the Opera* has a lognormal distribution looks clearly wrong. Kernel density estimates (KDEs) are superimposed on the histograms and are a more elegant way of showing this. KDEs, and the code required to obtain the figures, are discussed in Section 5.2 below.

[4] In `R` the desired number of bins will be modified to ensure a sensible choice of bin-width.

[5] Either because the histogram is incorrectly drawn – this is what was the default in Excel for many years – or because you are dealing with an incorrectly named 'bar chart', terminology best reserved for representations of discrete data.

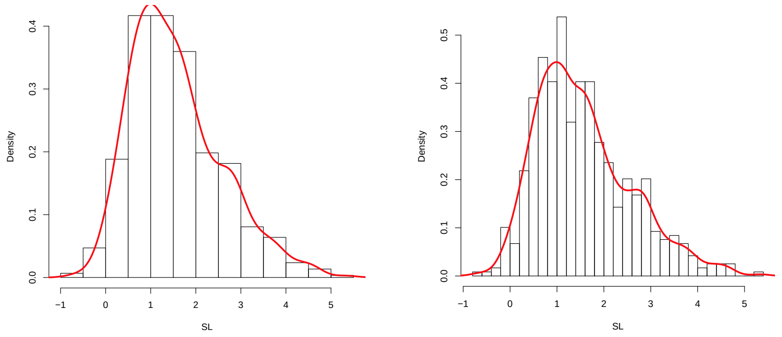

Figure 5.2: *Histograms of logged SLs for* A Night at the Opera. *See text for details.*

5.2 Kernel density estimates

A histogram is a particular form of *density estimate*. It has a number of limitations. One is that the appearance depends on the bin-width used and, the availability of 'rules' notwithstanding, the choice of what to present for publication purposes is 'subjective'. This is unavoidable, and true for other forms of density estimate, though the suspicion exists that software defaults are sometimes used without too much thought being given to the choice.

More seriously, and something avoided by other forms of density estimate, is that the appearance can be affected by the choice of anchor point. Histograms are also unwieldy for comparing the SL distributions of more than a small number of films. Of alternatives to the histogram only kernel density estimates (KDEs) are considered in detail here.

Statistical texts and papers that deal with KDEs can look mathematically forbidding, but the basic idea is very simple. One way of representing a single SL is as a point on a line between the maximum and minimum SLs. The point is replaced by a symmetric bump, centered on the SL. The spread of values covered by the bump are associated with different heights. This is done for each SL in turn. The KDE is then simply obtained by adding up the heights of the different bumps at each point along the line[6].

For *A Night at the Opera* the default KDE obtained using `plot(density(z))` is shown in Figure 5.3. This can be broken up into two commands, `NatOdens <- density(z)` followed by `plot(NatOdens)`[7].

In the same way that the appearance of a histogram is controlled by the choice of bin-width, the appearance of a KDE is controlled by the spread of the bumps used in its definition. This is determined by the *bandwidth* (or window width). The form of the bump (or *kernel*) can be chosen in various ways, but commonly a normal (or Gaussian) distribution is used, in which case the bandwidth depends on the standard deviation of the distribution. Large bandwidths give over-smoothed KDEs; smaller bandwidths give under-smoothed KDEs. As with histograms a variety of rules exist for the automatic choice of bandwidth and, as with histograms, it is usually sensible to experiment with different choices.

Looking at the default is usually useful in that it provides a starting point for such experimentation. In fact, because of the impact of the extreme point and the skewness with a long tail, the KDE in Figure 5.3 has the same problems for interpretation as the histogram.

If attention is confined to the region with SLs less than 40 a default bandwidth of 0.98 is obtained. Reducing this to 0.8 and tidying the graph up a bit for presentation using the following

[6] A technical issue is that because the bumps are spread around a point the KDE will have non-zero values lying outside the range of the observed SLs. There are ways of dealing with this, but it is simplest not to worry about it unless behaviour in the extreme tails is of especial interest.

[7] It is sometimes useful to do this, since `NatOdens` is an object containing information about the KDE than can be manipulated for other purposes. This is not pursued here.

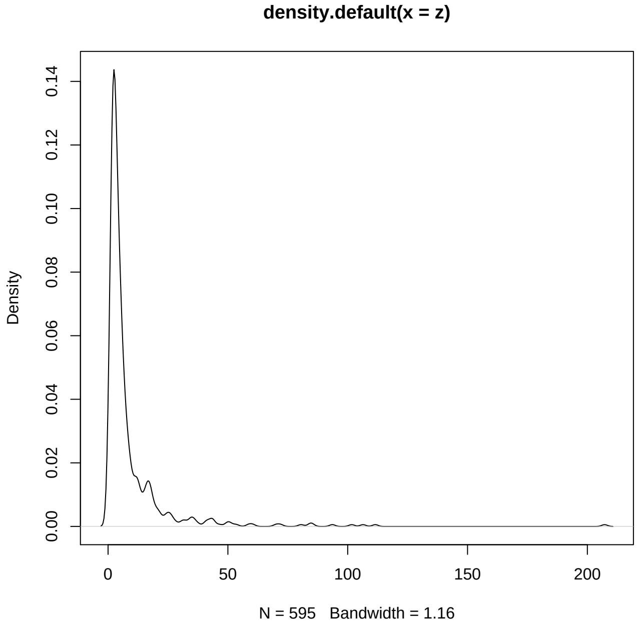

Figure 5.3: *The default KDE for the SLs of* A Night at the Opera.

code results in Figure 5.4. The main feature is that the tail of the distribution is somewhat bumpier than would be expected for a lognormal distribution

```
{\tt plot(density(z[z < 40], bw = .8), xlab = "SL (< 40), \\bandwidth = 0.8", main = "")}
```

This is a situation where looking at the SLs on a logarithmic scale is useful. Figure 5.5 illustrates the effect of different bandwidth choices; there is no need to select a subset of the data as done in previous figures. The code used was

```
plot(density(log(z)))
plot(density(log(z), bw = .20), xlab = "log(SL), bandwidth = 0.20", main = " ")
plot(density(log(z), bw = .12), xlab = "log(SL), bandwidth = 0.12", main = " ")
plot(density(log(z), bw = .35), xlab = "log(SL), bandwidth = 0.35", main = " ")
```

The upper-left plot is the default and used a bandwidth of 0.2447. With this as the starting point the upper-right figure reduces the bandwidth to 0.20. It doesn't make too much difference here; some of the features of the default are emphasised slightly more. The under-smoothed lower-left plot, with a bandwidth or 0.12 , shows too much spurious detail; the over-smoothed lower-right, with a bandwidth of 0.35, removes some of the potentially informative detail in the upper plots.

The clear skewness of all these plots is sufficient evidence of non-normality to demonstrate that the assertion that *A Night at the Opera* has a lognormal SL distribution is wrong. The upper plots (and under-smoothed plots) add further credence to this assertion by highlighting regions of 'lumpiness' of a kind inconsistent with a normal distribution.

If, for whatever reason, your preference is for histograms, it is easy enough to superimpose a KDE on a histogram. This was done in Figure 5.2, where the default KDE was superimposed on the default histogram, and the KDE with a bandwidth of 0.20 superimposed on the second histogram, using

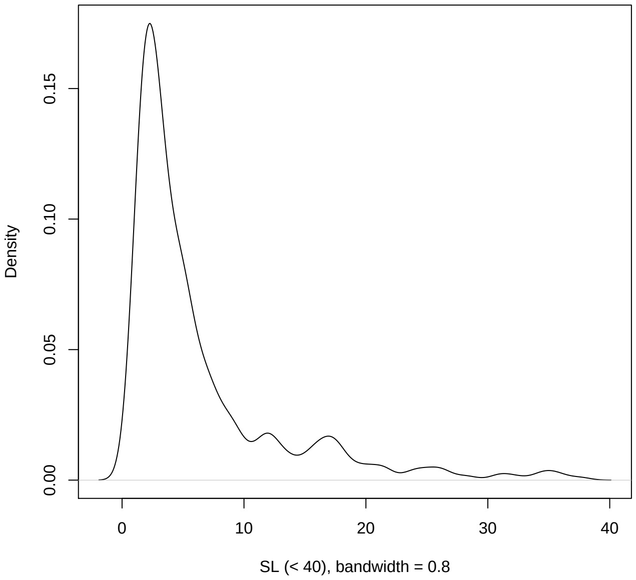

Figure 5.4: *The KDE for* A Night at the Opera*, using SLs less than 40 and subjectively chosen bandwidth of 0.8.*

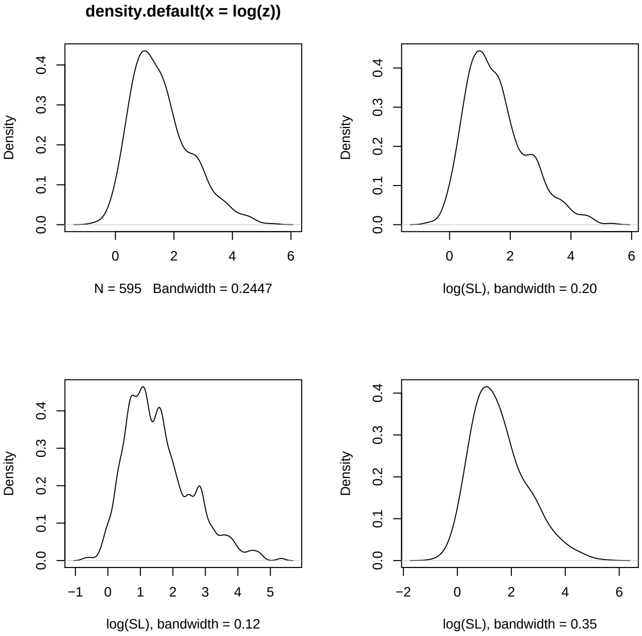

Figure 5.5: *The default KDE for the logged SLs of* A Night at the Opera *in the uopper-left, and KDEs for different choices of bandwidth.*

```
hist(log(z), freq = F, xlab = "SL", main = "")
lines(density(log(z)), lwd = 2, col = "red")

hist(log(z), breaks = 30, freq = F, xlab = "SL", main = "")
lines(density(log(z), bw = 0.2), lwd = 2, col = "red")
```

The arguments `lwd` and `col` control the line width and colour and are used here for presentational purposes. Though not used here `lty` can be used to control the line type; for example, `lty = 2` produces a dashed line.

5.3 Boxplots

5.3.1 Basics

The *boxplot* (or box-and-whisker plot) is a useful descriptive display. That for *A Night at the Opera*, shown in Figure 5.6, is obtained by `boxplot(z, horizontal = TRUE)` (the violin plot is discussed in the next section). The `horizontal = TRUE` argument produces the horizontal orientation of the plot preferred here; if a vertical orientation is desired it can be omitted, as this is the default.

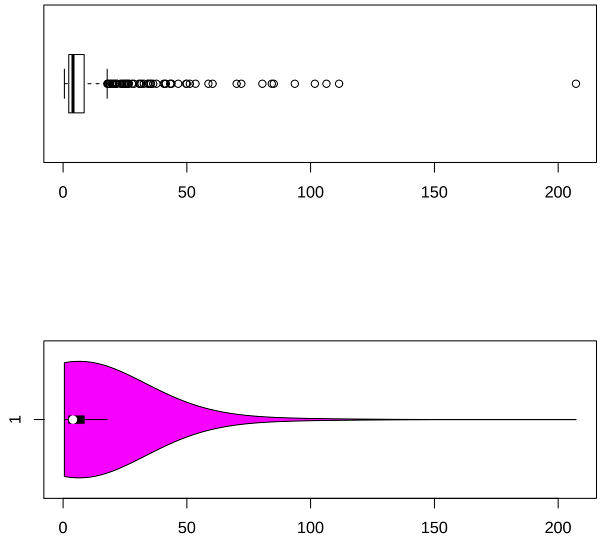

Figure 5.6: *A boxplot(upper) and violin plot (lower) for* A Night at the Opera*,*

The command `fivenum(z)` will produce the following 'five-number summary'

```
[1]   0.5   2.3   4.0   8.5 207.2
```

The upper and lower values are the minimum and maximum, which are the extremes of the boxplot. The second and fourth values are the lower and upper quartiles (Q_1 and Q_3), and define the limits of the box shown. The width of the box is IQR = ($Q_3 - Q_1$), where IQR is the *interquartile range*; the box thus contains the central 50% of the data. The middle value in the five-number summary, 4 in this instance, is the *median*, and is highlighted by the line within the box.

The 'whiskers', dashed lines that extend from the boxes, go by default as far as 1.5 × IQR from the limits of the box, beyond which 'unusual' points are shown individually. For typical SL data nothing 'unusual' will be highlighted in the left tail; for skewed distributions rather a lot of 'unusual' values may typically be shown in the right tail. Here the IQR is 6.2 (i.e. 8.5 - 2.3). Thus the upper whisker is broken at 17.8 (i.e. 8.5 + 1.5 × 6.2). There are 65 SLs greater than this, but only that at the maximum, 207.2, really stands out.

The definition of 'unusual' is arbitrary, with 1.5 as the default. It can be controlled by the `range` argument, and `range = 0` will extend the whiskers to the extremes. There has been some confusion in the cinemetrics literature about the identification of 'unusual' SLs, as defined above, with outliers. This, in turn, has fed into debates about the appropriate use of summary statistics for SL distributions. This is discussed further in Section 5.3.3.

5.3.2 Interpretation

The boxplot is a nice presentational device, particularly for comparative purposes, under certain conditions. Different software packages may present boxplots in different ways; for the control that can be exercised in `R` see `?boxplot`. Some software shows the mean as well as the median of the data; where 'unusual' values are indicated the default is often that used in `R`.

If a set of data is sampled from a symmetric unimodal (single-peaked) distribution (e.g., the normal) it should be approximately symmetrical, with the median near the middle of the box. Departures from symmetry in the form of skewness, can be manifested by the non-central placement of the median and whiskers of unequal length. For a reasonably symmetric underlying distribution the highlighting of 'unusual' values can draw attention to potential outliers; for very skewed distributions, typical of SLs, a lot of 'unusual' data may be highlighted that is, in fact, typical of what to expect from the underlying distribution.

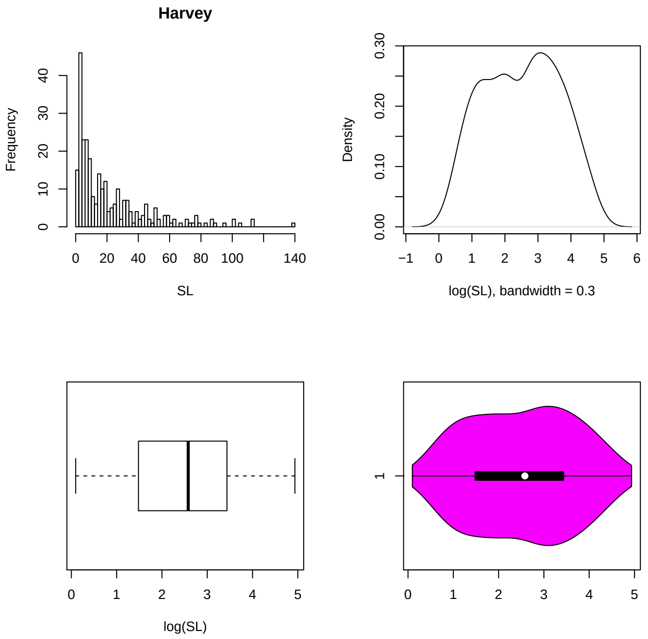

Figure 5.7: *A histogram for the SLs of* Harvey *and a KDE, boxplot and violin plot for the logged SLs of the film.*

The boxplot is not suitable for representing distributions with more than one mode, whether symmetric or not[8]. This is better illustrated using SLs after log-transformation, and this is done for *Harvey* (1950) in Figure 5.7.

[8] We will not worry here about the difference between 'major' and 'minor' modes.

I will leave it to readers to judge if they think the histogram for the untransformed data looks lognormal. *Harvey* has, in fact, one of the least lognormal distributions it is possible to find; looking at the KDE on a log-scale indicates one reason why – the central appearence is very different from that which is to be expected from normality, which is should be if lognormality obtains.

It is good practice to look at several kinds of plot rather than going straight to a boxplot; if you do the latter the near-perfect symmetry of the plot(bottom-left) may mislead you into thinking that the distribution is symmetric (and possibly normal). The problem is that the boxplot is constructed on the basis of the five-number summary only and takes no account of the density of the distribution, and hence cannot represent modes. The violin plot attempts to improve on this by showing the density as well as the boxplot. For the logged SLs of *Harvey* this produces a graph with a 'waist' indicating the absence of a clear single mode. For a symmetrical unimodal distribution it would be fattest in the middle, and tail off symmetrically on either side.

The R commands used for the figure were

```
hist(SL.Harvey, 50, xlab = "SL", main = "Harvey")

logH <- log(SL.Harvey)
plot(density(logH, bw = .3), xlab = "log(SL), bandwidth = 0.3", main = "")

boxplot(logH, horizontal = TRUE, xlab = "log(SL)")

library(vioplot)
vioplot(logH, horizontal = TRUE)
```

5.3.3 Boxplots and outliers

There has been discussion in the cinemetrics literature about the appropriate way to summarise SL distributions, in the context of stylistic analysis. The realtive merits of the median shot length (MSL) and average shot length (ASL) has attracted particular attention. A summary of the different positions on this topic is contained in debate in papers available on the Cinemetrics site at http://www.cinemetrics.lv/dev/on_statistics.php#

One argument against the use of the ASL, quite widely asserted, is that outliers are common in SL data, and that the ASL is sensitive to this (i.e. it is not 'robust'). Taken to extremes, this has led to suggestions that the ASL should not be used as a measure of 'film style', despite its widespread use for this purpose for over 30 years. My own position, elaborated in my contributions to the debate, is that the prevalence of outliers in SL data has been greatly exaggerated, as has their effect on ASL calculations where they do exist. It follows from this that the arguments against the ASL, based on the supposed effect of outliers, have little weight for practical purposes[9].

The opposite position is enunciated by Redfern in his contribution to the debate. The use of boxplots to demonstrate the prevalence of outliers is central to part of Redfern's argument, and one of the films he investigates, *The Scarlet Empress* (1934), is used as the peg on which to hang further discussion.

The histogram for the SLs is shown in Figure 5.8. The boxplot in the upper-right is the R default, and on the basis of it Redfern claims that there are 39 outliers. This is a serious misinterpretation.

The lower-left plot is for the logged SLs. A KDE or histogram of the logged data is unimodal, so use of the boxplot is justified. Discounting the two marginal outliers in the lower tail (induced by the log-transformation) there is no evidence of any unusual data (outliers). This is exactly what we would expect if the SLs followed a lognormal distribution, or at least a qualitatively similar skewed distribution. That is, the 39 supposed outliers in the boxplot for the untransformed data are no such thing; they are just what the boxplot is expected to show if the data are reasonably skewed with the features to be expected of a skewed distribution.

[9] Other arguments can be adduced; the focus here is only on outliers and the use of boxplots for identifying them.

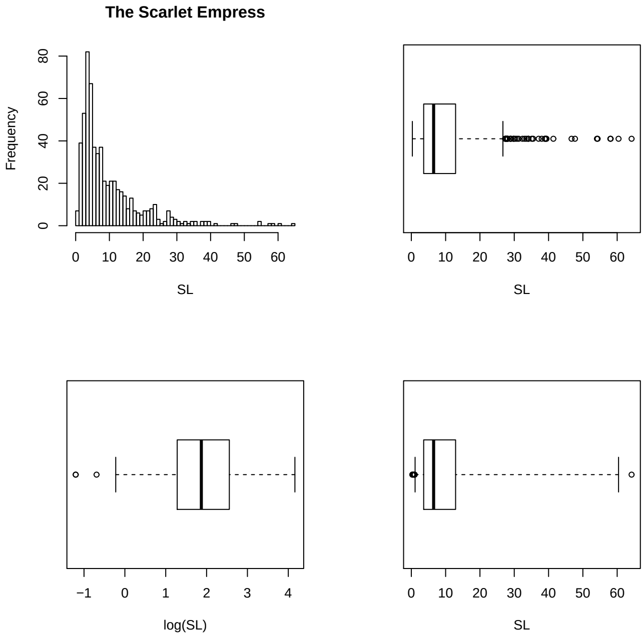

Figure 5.8: *A histogram for the SLs of* The Scarlet Empress *and a variety of boxplots – see the text for a discussion.*

In a research blog subsequent to the Cinemetrics debate, not posted there, Redfern (August 30th, 2012) (http://nickredfern.wordpress.com/category/statistics/) has conceded that his original use and interpretation of the boxplot, in terms of outlier detection, was inappropriate and presented an alternative version of the boxplot which, for *The Scarlet Empress*, gives rise to the final plot in Figure 5.8.[10] There is one very marginal unusual value in the right tail, so the original misinterpretation in terms of 39 outliers is clear.

Redfern's original, albeit flawed, argument has the merit of making it clear how an outlier was defined; this is not usually the case in publications where the prevalence of outliers in SL data is asserted. There is the suspicion that many commentators are bad at judging what is and isn't to be regarded as unusual in the tail of a lognormal distribution, and hence at judging what is an outlier[11]. A technical discussion on the ideas that underpin the final plot in Figure 5.8 may be useful in understanding why. The plot is obtained using

```
library(robustbase)
adjbox(SE, horizontal = TRUE, xlab = "SL")
```

remembering that the package `robustbase` has to be installed first (Section 3.3).

Redfern references Hubert and Vandervieren (2008) (HV) for his methodology; Googling brings up more than one earlier version of the paper. The limits beyond which unusual data are flagged in a boxplot are defined as

$$(Q_1 - k \times L), (Q_3 + k \times U)$$

where k is a constant and L and U are lower and upper values used in the definition. For both the default boxplot in `R` and that used in the `adjbox` command from the package `robustbase`,

[10] The blog uses *Lights of New York* for illustration; this does not affect the points made here.

[11] The lognormal assumption is convenient, but not essential here. Some authors who have asserted the prevalence of outliers do, however, also appear to accept the generality of lognormality. (e.g., DeLong *et al.* (2012); Cutting (http://www.cinemetrics.lv/cutting_on_salt.php)).

$k = 1.5$. For the standard boxplot $L = U = \text{IQR}$.[12]

What the HV paper does is to define an interval that is asymmetric about the median, using different values for L and U. These are defined after considerable experimentation, and are based on a robust measure of skewness that is one of several possibilities that might have been used. More experimentation with SL data would be useful, and it would be interesting to compare the results with those obtained using boxplots with log-transformed SLs. It is obvious that this ought to change perceptions about the extent of problems with 'outliers'. They do exist, the obvious one for *A Night at the Opera* being a case in point. Once identified in 'sensible profusion', or the lack thereof, their effect on ASL calculations – if this is of interest – can then be investigated, rather than it simply being asserted that they are problematic. As Salt (2012) suggests where outliers do exist it may be of more interest to examine the role they play in the 'style' of a film than to focus on numerical measures of 'style' that ignore them.

[12] The choice of k is arbitrary. For data sampled from a normal distribution with sample size 600, 3-4 unusual values are expected to be flagged using the standard boxplot. For *The Scarlet Empress* with $n = 601$ SLs this is much what is observed on the log-scale. Note that this does *not* imply that the SLs are lognormally distributed, though it is consistent with the possibility.