Chapter 4

Descriptive statistics

4.1 Introduction

This chapter provides a brief introduction to how some of the more commonly used descriptive statistics in cinemetrics can be calculated in R, with a section (Section 4.5) that illustrates what you might do with some of these. That section could be regarded as an continuation of the examples provided in Section 2, but also introduces some of the practicalities of using R, as well as showing how R encourages the exploratory analysis of data.

If interest centres solely on descriptive statistics commonly used in summarising SLs, such as the ASL or median, and the film is in the Cinemetrics database, they can be found there. The following Section 4.2 shows how these are obtained in R. It's easy enough to obtain the statistics for a large body of films in a single analysis, but it requires a different approach to the method of data entry described in the last chapter and assumed for the next few chapters on graphical analysis, so discussion of data entry is deferred. What you might do with statistics so acquired is, as just noted, explored in Section 4.5.

Section 4.3 is useful. Functions are collections of commands that you create and give a name to that, among other things, allows you to execute a range of different commands for a single film and/or repeat the same type of analysis for as many films as you wish. Once you've successfully created a function they save a lot of time if you are attempting more than infrequent 'once off' analyses. They can be made to do very complicated things, using complicated coding; nothing like that is attempted in these notes.

The other thing to know about at an early stage is what I've called data manipulation in Section 4.4. Just because someone, possibly yourself, has gone to the trouble of collecting data on SLs for a film doesn't mean you're obliged to use all of it. Provided you know what the shot type is, omission of those associated with the opening credits, or intertitles in silent films, might be of interest. The basics of this are discussed in the section.

The appropriate use of some of these apparently simple statistics, the ASL and median SL in particular, has generated a surprising amount of debate, as a search of the contributions to the Cinemetrics discussion board, or a glance at the Cinemetrics debate will show. Here is not the chapter for detailed discussion of the issues involved. They are, however, touched on in Section 4.5, which can also be regarded as an *hors d'oeuvres* for the graphically oriented chapters to follow.

4.2 Basics

Assuming the data for *A Night at the Opera* has been imported into R and SLs, in seconds, saved in `SL.NIGHT_at_the_Opera` it is convenient, to save typing, to define

```
z <- SL.Night_at_the_Opera
```

then

```
mean(z)           # ASL
median(z)         # MSL
median(z)/mean(z) # MSL/ASL
sum(z)/60         # LEN (length in minutes)
length(z)         # NoS (number of shots)
max(z)            # MAX
min(z)            # MIN
max(z) - min(z)   # Range
sd(z)             # StDev
sd(z)/mean(z      # CV (coefficient of variation)
```

gives the statistics reported with a Cinemetrics graph. The # character indicates a comment and this, and the information that follows – which are the names used in Cinemetrics – should not be typed.

Some of these statistics can be obtained with a single command. Thus, `summary(z)` returns

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.500   2.300   4.000   8.978   8.500 207.200
```

Here `1st Qu.` and `3rd Qu.` are the first and third *quartiles*, the difference between them being the *interquartile range*. They can be obtained using

```
quantile(z, .25)  # Q1
quantile(z, .75)  # Q3
IQR(z)            # IQR
```

the names being those given to the quantities when they are discussed later[1].

4.3 Functions

Googling the terms *descriptive statistics* and `R` brings up quite a few packages with functions that produce a range of statistics. None of these ever seem to do quite what you want, and it's tedious to keep typing the commands of the previous section if doing a lot of analyses.

It's easier to write a function to do exactly what you want. This sort of thing is often introduced at a later stage in introductory `R` notes, but worth engaging with early. Begin with something like

```
filmstats <- function(film) {}
filmstats <- edit(filmstats)
```

which brings up an edit screen where you can type in what you wish, thus ending with something like

```
function{film) {
z <- film
ASL <- mean(z)
MSL <- median(z)
ratio <- MSL/ASL
list(ASL = ASL, median = MSL, ratio = ratio) # List of statistics
}
```

listing just three statistics for brevity. Once done[2] `filmstats(SL.Night_at_the_Opera)` will return

[1] I'm resisting the idea of giving a precise definition of 'quartile'. There are nine different algorithms available in `R` which can give slightly different results that don't matter, given the size of data set usually available for analysis in cinemetrics.

[2] On exiting from `edit` mode and saving the work you will be notified of any errors. Sometimes these result in the work not being saved, which can be a pain. It is safest to copy work before you exit so it can be pasted back for correction if needed. Error messages are not always as helpful as you might wish.

```
$ASL
[1] 8.978319

$median
[1] 4

$ratio
[1] 0.4455177
```

As with much statistical software, numbers are reported with too many decimal places. This can be cured using `list(ASL = round(ASL,2), median = MSL, ratio = round(ratio,2))` in the function, rounding to two decimal places.

The obvious advantage of doing things in this way is that when invoking the function the data for any other film can replace `SL.Night_at_the_Opera`. It's possible to write functions to do the computations for several films in one go; this is left till later.

4.4 Data manipulation

An initial inspection of SL data may suggest either that analysis is better conducted after some omissions (e.g., opening credits), or that investigation of the effects of omitting some shots (e.g., suspected outliers) is desirable. This is straightforward if you know what you want to omit. With no particular film in mind, and for example

```
mean(z[-1])            # Omit the first shot
mean(z[-c(1:4)])       # Omit the first 4 shots
mean(z[-347])          # Omit shot 347
mean(z[-c(5,108,347)]  # Omit the three shots listed
```

The 69th SL for *A Night at the Opera*, at 207.2, is somewhat longer than the next longest at 111.5. The command `filmstats(SL.Night_at_the_Opera[-69])` returns an ASL of 8.64 compared to 8.98 using all the data.

The command `z <- sort(z, decreasing = T)` will reorder the SLs in `z` from largest to smallest. If the effect of possible outliers on, for example, ASL calculations is a concern this can be investigated systematically basing analyses on `z, z[-c(1)], z[-c(1:2]), z[-c(1:3)]` etc.

Suppose that a film has been recorded in advanced mode, as discussed in Section 3.2 for *Lights of New York*, where the type of shot is recorded, one of the categories being titles, `exp.tit`. The SL and type variables were named `SL.LONY` and `SL.LONY.Type` in `R`. Suppose an analysis is to be undertaken omitting title shots. The necessary data is most simply created using

```
SL.LONY.NoTitle <- SL.LONY[SL.LONY.Type != "exp.tit"]
```

The `!=` part of the command is to be read as 'not equal to' so what is being selected are shots that are *not* categorised as titles, `exp.tit`, in the type variable[3]. If you wanted to select titles only use `==` rather than `!=`.

4.5 Illustrative graphical analyses

There is, as the Cinemetrics debate shows, some argument about the appropriate choice of statistics for summarising SL distributions, most obviously whether the ASL or median SL is a better descriptor of 'film style'. A similar, if less publicised, choice exists in terms of choosing a measure of dispersion, but usually the standard deviation goes with the ASL and the interquartile range

[3]You need to be careful to both enclose the type category selected in double inverted commas and use square rather than round brackets. This can be tedious and it's easy to slip-up, but you get used to it.

(IQR) with the median. Redfern (2010a) has noted some alternatives to the IQR, all robust measures of dispersion (or scale), considered below.

Rather than arguing from principle about what should be the preferred choice, a pragmatic view is to see if the choice makes much difference when a large body of films are considered together. Those used here are 134 of the 150 described in Cutting *et al.* (2010). These are (mostly) Hollywood films from 1935-2005, selected as 'high-profile' in some way, at least in their day; 16 of the 150 are not used because they have recorded zero or negative SLs that are presumably recording errors (Redfern, 2012a) which preclude the analysis after log-transformation carried out elsewhere in these notes.

Four variables have been selected for illustration, the ASL, median SL, standard deviation (SD) and IQR. Figure 4.1 shows examples of what are called pairs plots or scatterplot matrices – just a collection of all the two-way plots it is possible to get for the variables. Each pair is plotted twice, with the axes reversed. Thus the first plot along at the top is of ASL against median SL; the first one down on the left is of median SL against ASL etc. It is assumed that four variables

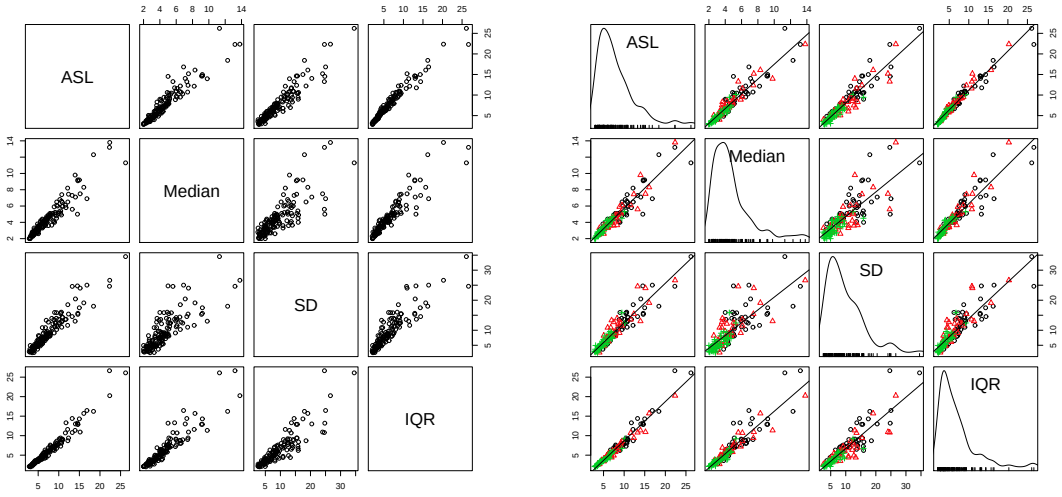

Figure 4.1: *Scatterplot matrices for selected variables .*

have been created for use in `R` named `ASL`, `Median`, `SD` and `IQR`. This is assumed here for ease of exposition[4].

The two plots shown are, enhancements apart, the same. The basic plot to the right is obtained from

```
stats1 <- cbind(ASL, Median, SD, IQR)
pairs(stats1)
```

Some broad features are immediately obvious. One is that there is a reasonably good linear relationship between most pairs of variables. The strongest, interestingly, is that between the ASL and IQR, the weakest that between the SD and IQR. One implication of this, with caveats to be entered, is that for any study that looks for patterns in a large body of data, it probably doesn't matter much whether the ASL or median SL is used as a measure of location. Baxter (2012a) has suggested this and, equivalently, Salt (2011) has noted 'a relatively fixed ratio between the ASL and Median', with 82% of films in his sample of 1520 having a Median/ ASL ratio in the range 0.5-0.7, a 'remarkable fact' that 'demands explanation'.

[4]This is actually what I did here, operating on data for individual films already entered into `R`. If the data are created externally, in the form of a table put together and published by someone else, for example, and if they are in Excel, the table can be read in and individual variables extracted and named as described in Section 3.2.2.

The first caveat is that for films with the larger ASLs (greater than about 12) and median SLs (greater than about 7-8) the linear relationships are not so strong. Salt has suggested in a number of places that general patterns can be expected to break down for films with large enough ASLs, 15 seconds sometimes being suggested as a rule of thumb. The second caveat is that there is a significant minority of films for which none of these statistics are particularly suitable summary measures (Baxter, 2012a), and these may or may not show up as departures from the general trends on the plots.

As well as noticing broad trends, and general departures from them, one can begin to pick out films that look statistically unusual for some reason. For example, three or four films stand out in the top-right corner of those plots involving the ASL and median. These are, chronologically, *Detour* (1945), *Harvey* (1950), *The Seven Year Itch* (1955) and *Exodus* (1960). Detailed examination of these, of the kind discussed in Section 6.4.4, suggests that *Harvey* and *The Seven Year Itch* are not suitably summarised by the ASL or the median SL, whereas the other two films are among the few that Redfern (2012a) (using the same sample as here) finds to have acceptably lognormal distributions[5].

It's noticeable, also that these films are relatively early in the 1935-2005 date range. No analysis takes place in a vacuum, and it's well-known that ASLs (and median SLs) have declined steadily over the last 50 years and more. This is exploited in the left-hand panel of Figure 4.1 where films are coded into three date ranges, 1935-1955 (black open circles), 1960-1975 (red open triangles) and 1980-2005 (green crosses)[6]. The relationship between date and size of the ASL and median is immediately apparent, with the later films concentrated in the lower left for each plot and showing less obvious departure from a linear scatter.

The date codes were (1, 2, 3) and a variable `DateGroup` was created for this in `R`. The plot was produced using the `scatterplotMatrix` function from the `car` package, which needs to be installed (Section 3.3). Thus

```
library(car)

scatterplotMatrix(~ ASL + Median + SD + IQR, smooth = F, by.groups = F,
groups = DateGroup, legend.plot = F)
```

does what's required. This is quite a useful function so worth discussing in a little detail. The variables you want plotting are listed here separately, preceded by $\sim$. By default a straight line is fitted through the data, which can be suppressed but does no harm here and is left in. A smooth (loess) curve is also fitted through the data by default (Section 7.5), but this seems superfluous here, given the linearity, and has been suppressed by `smooth = F`. The `groups = DateGroup` argument tells `R` where to get the labelling information from; lines are fitted through the separate groups by default, which may be useful or may complicate the plot too much and has been suppressed here by `by.groups = F`. A legend is supplied by default, which can sometimes obscure things, so has been omitted here by `legend.plot = F`.

By default kernel density estimates (KDEs) (Section 5.2) are shown on the diagonal, for each variable separately, with a 'rug' showing individual data points. This is optional and can be replaced by other graphical displays, or omitted (use `?scatterplotMatrix` to see the options). Here, they're harmless and serve mainly to show the similarly skewed nature of the variables.

One of the great strengths of `R` is the ease with which rapid data exploration can be undertaken including experimentation with different date groups. What's eventually selected for presentation, if that is the end result, is usually arrived at iteratively. For many of the illustrations in these notes, tidying up figures for presentation took longer than the few minutes needed for the several analyses needed before a choice was made. That is, for exploratory analysis, `R` is very quick once the data are set up.

[5] Redfern's criteria for judging lognormality are too stringent, but it means that any film accepted as lognormal by his methods is likely to be accepted as such by other approaches. This comment is not, by the way, opinion. It is easy to show that, hypothetically, two films with identical SL distributions, differing only in the number of shots, can be judged differently, one as lognormal and one as not, by Redfern's methods.

[6] Cutting *et al.* (2010) selected films at five year intervals.

Figure 4.1 is perhaps on too small a scale to see fine detail but broad patterns are evident. On screen, and depending on its size, the complexiy of any pattern and your eyesight, up to 8-10 variables can be examined usefully at one go, it always being possible to separately 'home-in' on any plots of particular interest. It's also possible to concentrate on particular areas of interest by 'magnifying' the plot. Figure 4.2 illustrates, where only films with an ASL of less than 7.6 seconds are shown. This was designed to include all the films from 1980 on, with the exception of *Coal Miner's Daughter* (1980) and *Cast Away* (2000) whose ASLs, 10.1 and 9.5, are somewhat bigger than the next largest of 7.5 seconds.

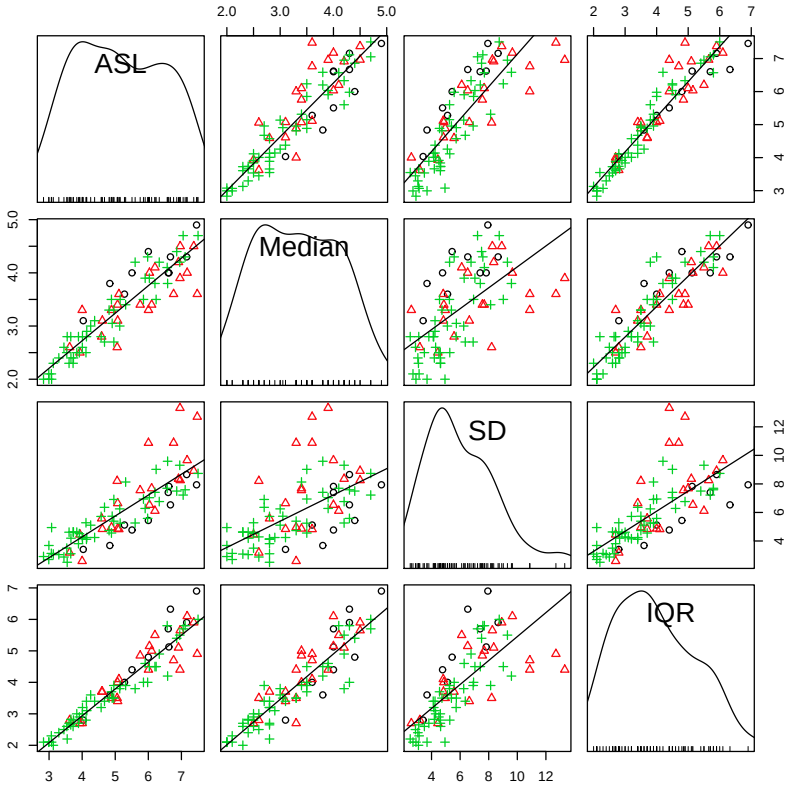

Figure 4.2: *A 'magnified' version of the right-hand panel of Figure 4.1 for films with ASLs less than 7.6 seconds.*

The patterns previously noted 'hold-up' pretty well – there's always the danger that interesting detail can be obscured by inclusion of the more extreme values in a data set (now omitted) which can 'squash-up' the smaller values to the point that detail is lost[7]. Not evident from the previous figure, and standing out a little, are four films with SDs greater than 10 that are larger than expected given other pattern, and in relation to the ASLs and medians. These are all from the 1960-1975 period and are *Those Magnificent Men in Their Flying Machines* (1965), *Five Easy Pieces* (1970), *Jaws* (1975) and *The Rocky Horror Picture Show* (1975). The first and last of these have one obvious and extreme outlier and *Jaws* has, perhaps, three less obviously extreme. *Five Easy Pieces* has what I would characterise as an unusual right tail, rather than a small number of obvious outliers.

The intriguing consistency of the relationship between the ASL and IQR remains, but is not pursued here. It is however, a useful peg on which to hang a discussion of measures of dispersion.

[7] One of the reasons for sometimes using logarithms, incidentally, which can remove this problem.

Redfern (2010a) has discussed a number of robust alternatives to the SD, in addition to the IQR. One is the median absolute deviation (MAD) defined as

$$\text{MAD} = \text{med}_i|x_i - \text{med}_j x_j|$$

or, in words (a) calculate the median of all the SLs; (b) difference each SL from this median; (c) calculate the median of these differences.

Following Rousseeuw and Croux (1993), whose notation is used here, the other two are S_n and Q_n. The first of these is defined as

$$S_n = \text{med}_i(\text{med}_j|x_i - x_j|)$$

where, in words, (a) for the first shot calculate the difference between its SL and that of each of the other shots; (b) find the median of these differences; (c) repeat this process for each shot in turn so you end up with n medians; (d) find the median of these n medians.

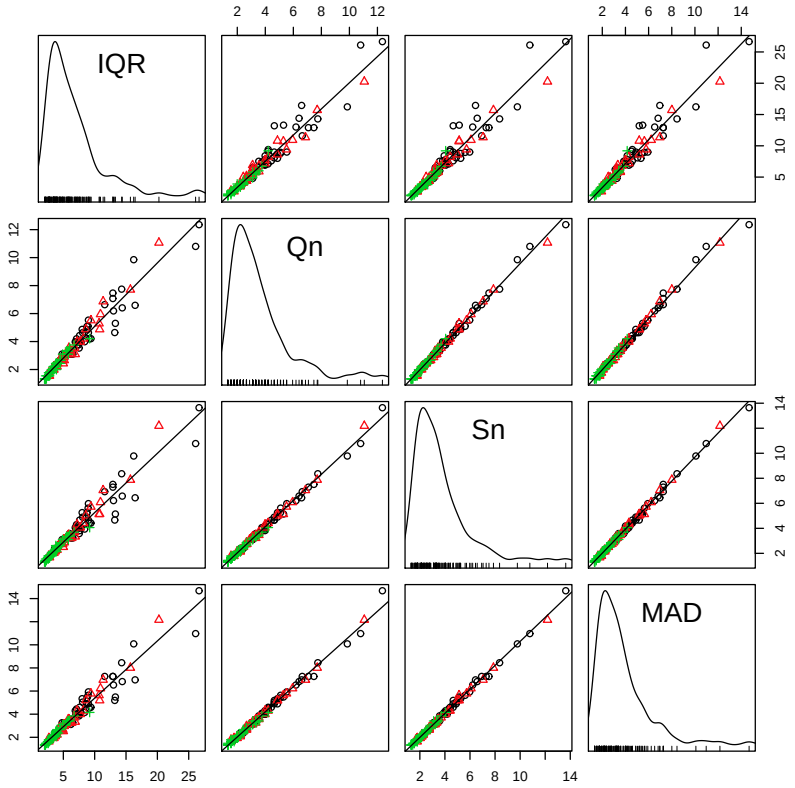

Figure 4.3: *A scatterplot matrix for four robust measures of dispersion descriptive of 134 Hollywood films (1935-2005).*

For large samples Q_n is defined as the first quartile of

$$k_q(|x_i - x_j|; i < j)$$

which are just the absolute differences in SLs between all possible pairs of shots. In these definitions k_s and k_q are constants that can be ignored for present purposes[8].

For the SLs of an individual film, `z` say, the statistics are easily obtained as

[8]The constants given in Rousseeuw and Croux (1993) are designed to achieve consistency, meaning that for a very large sample from a normal distribution you get pretty much the same values of S_n and Q_n as for the estimate of the SD. They don't matter here because only their relative values are of interest.

```
mad(z)
library(robustbase)
Sn(z)
Qn(z)
```

the package, `robustbase`, having been installed to obtain the last two statistics. How these can be obtained for a body of films in a single analysis is discussed in a later chapter.

The theory that establishes the properties of the statistics is quite complex, but the idea behind them is simple enough. The statistics differ in their 'degree of robustness', with the IQR at the bottom, MAD in the middle and the other two at the top. For practical purposes, and typical SL data, do these theoretical properties much matter? Figure 4.3 suggests not.

The results for MAD, S_n and Q_n are almost indistinguishable. There is some difference in results for the IQR, most obviously at the larger values of the statistics. This fairly closely mirrors what was the case for the ASL and median in Figure 4.1. The impression left by this last figure, and Figure 4.2 is that any discrepancy in conclusions that might be drawn, would arise in choosing between the SD and IQR (or other robust measures of dispersion).

The general impression, though it merits more systematic investigation, is that for later films from about the mid-1970s, and for comparing a large number of films, the choice of summary statistics to use may not matter much. The same patterns emerge. For earlier films there is sufficient evidence (Baxter, 2012a) that a significant minority exhibit characteristics that render them unsuitable for summary using just one or two statistics, so that generalizations about pattern based on just the ASL and SD, or median and IQR, should be approached with caution.

Individual films are the building blocks used when constructing pattern, and close inspection of them means that some may need to be rejected for the purposes of erecting structure. It is arguable, if not undeniable, that in looking at individual films, or comparing small numbers of films, a much richer analysis is possible using graphical rather than summary descriptive methods of analysis. Graphical methods are the subject of the next few chapters.